

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

Odd O. Aalen

*Department of Biostatistics, Institute of Basic Medical Sciences,  
University of Oslo, Oslo, Norway  
E-mail: o.o.aalen@medisin.uio.no*

Richard J. Cook

*Department of Statistics and Actuarial Science,  
University of Waterloo, Waterloo, ON, N2L 3G1, Canada  
E-mail: rjcook@uwaterloo.ca*

Kjetil Røysland

*Department of Biostatistics, Institute of Basic Medical Sciences,  
University of Oslo, Oslo, Norway  
E-mail: kjetil.roysland@medisin.uio.no*

## Summary

Statistical methods for survival analysis play a central role in the assessment of treatment effects in randomized clinical trials in cardiovascular disease, cancer, and many other fields. The most common approach to analysis involves fitting a Cox regression model including a treatment indicator, and basing inference on the large sample properties of the regression coefficient estimator. 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 problem arises because the risk sets beyond the first event time are comprised of the subset of individuals who have not previously failed. The balance in the distribution of potential confounders between treatment arms is lost by this implicit conditioning, whether or not censoring is present. Thus while the Cox model may be used as a basis for valid tests of the null hypotheses of no treatment effect if robust variance estimates are used, modeling frameworks more compatible with causal reasoning may be preferable in general for estimation.

*Keywords:* causation, collapsible model, confounding, hazard function, survival data

## 1 INTRODUCTION

Kaplan-Meier estimation (Kaplan and Meier, 1958) and Cox regression models (Cox, 1972) play a central role in the assessment of treatment effects in randomized clinical trials of cardiovascular disease, cancer and many other medical fields. When interest lies in delaying the time to an undesirable event, both methods naturally accommodate right censoring. The Cox regression model, however, is particularly appealing because it yields a simple summary of the treatment effect in terms of the

hazard ratio, which can be interpreted as a type of relative risk. The semiparametric nature of the model, the ability to stratify in multi-center trials, and the connection with the log-rank test have also contributed to its widespread use.

Randomization is usually viewed as the ideal approach for eliminating the effects of confounding variables to ensure causal inferences can be drawn. In randomized clinical trials with a survival outcome, it is customary to express treatment effects by estimates of the hazard ratio often without adjusting for even known prognostic variables. When viewing Cox regression from a causal point of view, omission of these terms creates problems which do not seem to be widely appreciated in some fields.

First we would like to point out that several limitations of the Cox model are well known. The selection effects well documented in frailty theory (e.g. Aalen et al., 2008), constitute one example to be mentioned below. A related issue is the fact that Cox models which condition on different sets of covariates cannot simultaneously be valid (Ford et al., 1995). Our intent is to focus on the limitations of the Cox model from a causal point of view, some of which have been pointed out in the epidemiological literature (Greenland, 1996, Hernán, 2010, Hernán et al., 2004; Hernán and Robins, 2015, Section 8.3). Our paper is partly a review of these ideas, but we discuss this in a broader setting and wish to present the issues to a biostatistical audience concerned with survival analysis.

The basic notion may be explained intuitively as follows: for a randomized study of survival times the first contribution to the Cox partial likelihood is based on a randomized comparison, but subsequent partial likelihood contributions are based on biased comparisons. This bias arises when there are known or unknown factors influencing survival which are not controlled for in the analysis; we believe this is almost always the case.

Our aims are to further clarify these issues through a simple mathematical argument and to provide further discussion, including the relationship to frailty theory. Cox regression is used routinely in clinical trials for the analysis of time to event response. We suggest that analysts applying these methods take it for granted that the hazard ratio gives a valid representation of the causal effect. It is therefore important to realize the limitations of this analysis. We focus here on settings where the goal is to assess the effect of treatment versus a control intervention on the time to an event in the context of a randomized clinical trial.

The theory of counting processes plays a fundamental role in survival and event history analysis. The probabilistic structure of counting processes is defined in terms of intensity processes. If  $N(t)$  is the process counting events over time (e.g. occurrences of undesirable clinical events), then the intensity process is :

$$\lambda(t; \mathcal{H}(t)) = \lim_{\Delta \rightarrow 0} \frac{1}{\Delta} P(N(t + \Delta) - N(t) = 1 | \mathcal{H}(t)) \quad (1)$$

where  $\mathcal{H}(t)$  is the history of the event process observed up to time  $t$ ; we assume that the information from the past grows as  $t$  increases. In words, the intensity process gives the probability of an event occurring in a small time interval  $[t, t + \Delta)$  given the past history observed to an instant before  $t$ . A precise mathematical treatment requires tools from martingale theory (Aalen et al., 2008).

The intensity process is based on the *concept of prediction* in that it involves the exploitation of information unfolding over time in order to model the instantaneous risk of an event. The utility of intensity-based modeling is clear in settings where interest lies in understanding dynamic features of a complex event process and associated risk factors. An important question in the setting of a clinical trial, however, is whether the event intensity can also be given a structural, or causal interpretation. Generally, of course, one cannot expect a structural interpretation due to unmeasured confounders or other issues.

Pearl (2009) gives a careful discussion of the requirements for a model to admit causal inferences and introduces the important tool of the *do*-operator. In the present context a causal statement can be made with respect to a particular intervention  $E$  if

$$\lambda(t|\mathcal{H}(t), e) = \lambda(t|\mathcal{H}(t), do(e)) \quad (2)$$

where  $do(e)$  means that the value of  $E$  is set to  $e$  by an external manipulation. The intervention  $E$  may correspond to changing an element in the history of the process, and the intensity process clearly does not tell you what the result of that would be unless (2) is fulfilled.

In the analysis presented here, the concepts of collider effect and controlled direct effect also play a central role, and we shall follow the ideas of Pearl (2009) here as well. In a process setting one has to show some care in applying these ideas (Aalen et al., 2014).

## 2 RANDOMIZED COMPARISON OF TREATMENTS IN SURVIVAL ANALYSIS

### 2.1 CONFOUNDING AMONG SURVIVORS

The risk of an event may vary considerably between individuals under study due to known and unknown factors. Examples of known factors might be the duration and stage of the disease at the time of recruitment, smoking status, and so on. Other features such as dietary history, level of physical activity, and environmental exposures are more difficult to measure and are typically not recorded. Genetic and epigenetic features are likely to have strong effects. We consider and represent these various influential factors collectively in a variable  $Z$ . Note that this variable is not routinely considered to have a confounding effect *per se* in a clinical trial with randomization, since randomization ensures  $Z \perp\!\!\!\perp X$  at the time of treatment assignment. Randomization does not, however, eliminate the effect of this variable on the outcome and as a consequence does not ensure the same independence among individuals surviving to any time  $t > 0$ , i.e. we do not necessarily have  $Z \perp\!\!\!\perp X|T > t$ .

To expand on this, we assume that the hazard rate for an individual in a population of interest is given as  $h(t, X, Z)$ . Here  $X$  is a treatment indicator and  $Z$  is the variable that contains individual specific quantities that influence survival. If the treatment assignment is randomized at  $t = 0$ ,  $X$  and  $Z$  are stochastically independent at that time. For simplicity, we shall assume here that  $X$  and  $Z$  are discrete, but a similar argument would of course hold for continuous variables. The joint distribution of  $X$  and  $Z$  for survivors to time  $t$  is given by:

$$\begin{aligned} P(Z = z, X = x|T \geq t) &= \frac{P(Z = z, X = x, T \geq t)}{P(T \geq t)} \\ &= \frac{P(T \geq t|Z = z, X = x)P(Z = z, X = x)}{P(T \geq t)} \\ &= \frac{\exp(-\int_0^t h(s, x, z)ds)P(Z = z)P(X = x)}{P(T \geq t)}. \end{aligned} \quad (3)$$

This expression can only be factored with respect to the two covariates for all  $t$  if  $h(t, x, z)$  is additive in the following sense:

$$h(t, x, z) = a(t, x) + b(t, z) \quad (4)$$

If this is not the case, then it is apparent from (3) that  $X$  and  $Z$  are not independent among the survivors (i.e. given  $T \geq t$ ).

Note that the hazard of the Cox model does not satisfy the additivity assumption in equation (4). This is also the case for a marginal Cox model containing only the covariate  $X$ : if the joint hazard in  $X$  and  $Z$  satisfies (4) then the marginal model for  $X$  would retain an additive component and would not be of the Cox form.

The lack of independence between  $X$  and  $Z$  among the survivors to time  $t$  has an important implication for the Cox model since the partial likelihood is built up by considering contributions which condition on survival to increasingly large times. Despite randomization at  $t = 0$ , the contributions to the Cox partial likelihood following the first failure will not be in a randomized setting. 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.

The effect we are discussing depends on the probability of surviving; as shown in (3). If the event in question is rare, then the effect will be small.

## 2.2 AN ILLUSTRATIVE CALCULATION

Suppose  $X$  and  $Z$  are binary 0–1 variables with  $P(Z = 1) = \pi$  and where  $X$  is assigned by balanced randomized with  $P(X = 1) = 0.5$ . Under a Weibull proportional hazards model we define  $H_0(t) = (h_0 t)^\kappa$  and consider the survivor functions  $P(T \geq t|X, Z) = \exp(-H_0(t) \exp(\beta_1 X + \beta_2 Z))$ ,  $P(T \geq t|X) = E_Z\{\exp(-H_0(t) \exp(\beta_1 X + \beta_2 Z))\}$ , and  $P(T \geq t) = E_X[E_Z\{\exp(-H_0(t) \exp(\beta_1 X + \beta_2 Z))\}]$ . We set  $h_0 = 1$  and  $\kappa = 1$  and consider  $\beta_1 = \log 0.5$  to correspond to a strong treatment effect and  $\beta_2 = \log 4$  to reflect a highly influential risk factor. We consider the  $p$ th percentile  $Q_p$ , of the marginal survival distribution, satisfying  $1 - p = P(T > Q_p)$ , with  $p = 0, 0.10, 0.25, 0.50, 0.75$  and  $0.90$ . In Figure 1 we display  $P(Z = 1|X, T \geq Q_p)$  from study entry for the case of common risk factor ( $\pi = 0.50$ ). There is a striking imbalance in the risk factor evident in the second and third quartiles of  $T$ . Evidence of treatment effect from individuals at risk at this time is therefore heavily influenced by this risk factor. The light grey line reflects the log odds ratio characterizing the association between  $Z$  and  $X$  given  $T > Q_p$ , clearly conveying the evolving dependence between treatment and the risk factor.

The Weibull proportional hazards model is the only parametric model which can be reformulated as a location-scale model (Cox and Oakes, 1984). We can therefore write it equivalently as

$$Y = \gamma_0 + \gamma_1 X + \gamma_2 Z + \tau W$$

where  $Y = \log T$ ,  $\tau = \kappa^{-1}$  is a dispersion parameter, and  $W \perp\!\!\!\perp (X, Z)$  has a standard extreme value error distribution; note also that  $\gamma_j = -\tau\beta_j$ ,  $j = 1, 2$  and  $\gamma_0 = -\log h_0$ . While we may think of this model as operating in the population of interest, when the treatment assignment is randomized in a clinical trial, we aim to estimate features of the distribution of  $T|X$ . The effect of interest is contained in the systematic part of the model for  $Y|X$ ,

$$E(Y|X) = \gamma_0 + \gamma_1 X + \gamma_2 E(Z|X) + \tau E(W|X) = \gamma_0^* + \gamma_1 X .$$

As a consequence, the effect of treatment in the location scale formulation, or its alternative utilization in terms of an accelerated failure time model (Wei, 1992), does yield a causal measure of treatment effect. The collapsibilities of this model and the causal interpretation have led to the widespread use of accelerated failure time models in causal inference (e.g. Robins, 1992). Consistent estimation of  $\gamma_1$  based on the model for  $T|X$  will require correct specification of the error distribution, but estimation of the coefficients in the location scale formulation is quite robust to misspecification of the error distribution (Gould and Lawless, 1988; Lawless, 2003, Section 6.3.4) and fitting semiparametric versions of this model can provide robust estimation under some forms of censoring.

We consider a brief simulation study to illustrate the utility of the location-scale model in this setting. We consider  $\pi = 0.10$  and  $0.50$  where  $X$  is determined by randomization; so  $P(X = 1) = 0.5$  and  $X \perp\!\!\!\perp Z$ . An administrative censoring time  $C^\dagger$  was set such that  $P(T > C^\dagger) = 0.10$  and an exponential random censoring time accommodated early withdrawal with rate  $\rho$  selected to give a net censoring rate of 30% or 50% respectively. Analyses were carried out based on Cox regression and

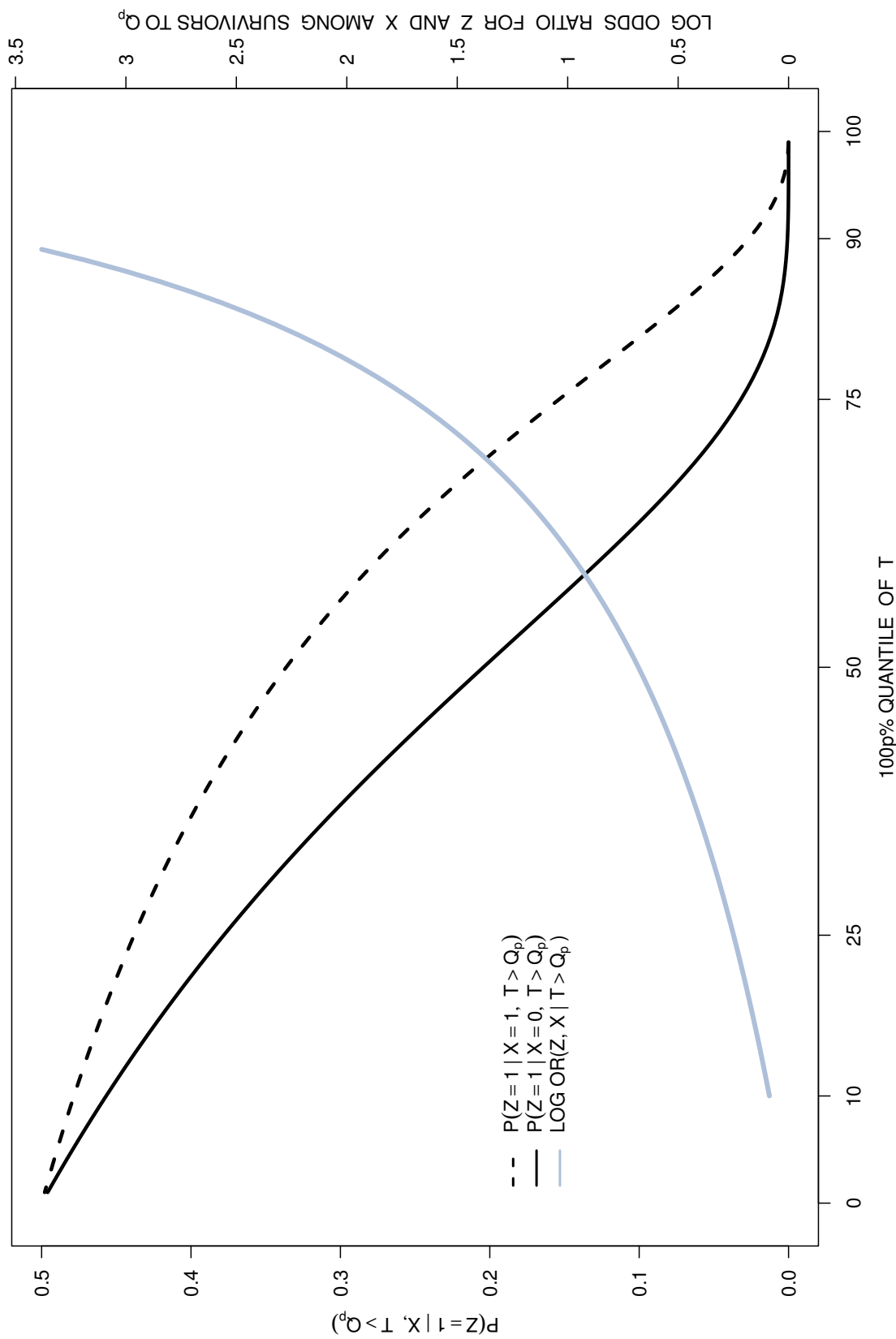


Figure 1: Conditional distribution of binary risk factor ( $P(Z = 1|X = x, T \geq Q_p)$ ) where  $Q_p$  is the  $p$ th percentile of the marginal distribution of  $T$ ) arising from a Weibull regression model with  $h_0 = 1$ ,  $\kappa = 1$ ,  $\beta_1 = \log 0.5$  and  $\beta_2 = \log 4$ ;  $X = 0$  for control group and  $X = 1$  for treated group;  $P(Z = 1) = \pi = 0.5$ ; also displayed is the odds ratio for  $Z$  and  $X$  given survival to  $Q_p$ .

a semiparametric accelerated failure time model using the method of Brown and Wang (2005): point estimates and robust variance estimates were recorded for each simulated sample.

With  $\kappa = 1$ , the regression coefficient in the accelerated failure time model is simply the negative of the coefficient in the Cox model and the empirical biases and coverage probabilities are evaluated relative to the respective true values.

Table 1: Empirical properties of estimators obtained by fitting marginal Cox and semiparametric accelerated failure time models for  $T|X$  when the correct model is a Weibull proportional hazards/accelerated failure time model for  $T|X, Z$ ;  $P(Z = 1) = \pi$ , 10% administrative censoring and CEN% reflects net censoring incorporating random withdrawal, 500 individuals per dataset; nsim=2000

$\pi$	CEN%	Cox Model				AFT Model			
		EBIAS	ESE	RSE	ECP%	EBIAS	ESE	RSE	ECP%
0.1	30	0.043	0.111	0.109	92.2	0.003	0.135	0.131	94.5
	50	0.048	0.129	0.130	92.6	0.004	0.161	0.157	94.3
0.5	30	0.147	0.108	0.108	72.9	0.001	0.146	0.142	93.8
	50	0.120	0.131	0.129	84.3	0.002	0.169	0.163	93.8

EBIAS is mean estimate minus  $\beta_1$  (Cox) and  $\gamma_1$  (AFT) respectively, ESE is the empirical standard error, RSE is the mean robust standard error, and ECP% is the empirical coverage probability

The bias evident in the estimation of  $\beta_1$  from the Cox model arises because it is not collapsible and is sensitive to the distribution of the prognostic factor as well as the censoring distribution; this is the case with misspecified Cox models generally, for which the large sample properties of associated estimators are now well known (Struthers and Kalbfleisch, 1986). Hypothesis tests directed at detecting treatment effects are valid in such settings, however, provided robust variance estimates are used (Lin and Wei, 1989). The empirical biases in the estimators of the coefficients in the accelerated failure time model are negligible in all cases and so are insensitive to the distribution of the covariate and censoring rates, and there is generally close agreement between the empirical and mean robust standard errors.

### 3 INTERPRETATION IN A CAUSAL INFERENCE SETTING

The concept of a collider is used to clarify the effects of selection bias in causal reasoning. A collider is present in a directed acyclic graph (DAG) if two arrows meet at a node. When conditioning on a collider the effects may “pass through” the collider, and since it does not exert a causal effect, a bias may be induced in the estimate of the effect of the intervention. See Pearl (2009) for a general introduction to these matters.

Consider two times,  $t$  and  $t + \Delta$  where  $\Delta > 0$ , and let  $S_t$  and  $S_{t+\Delta}$  indicate survival up to these times respectively. A model for the causal structure of this situation is shown in the directed acyclic graph of Figure 2.

If we consider the comparison of treatments with respect to  $S_t$  or  $S_{t+\Delta}$  separately, then a random allocation will ensure a valid assessment of the causal effect. For instance, when considering the effect of  $X$  on  $S_{t+\Delta}$ , the noncausal path, via  $Z$ , is closed as long as we do not condition on  $S_t$ . The node  $S_t$  is a collider and closes the path via  $Z$  as long as it is not conditioned upon.

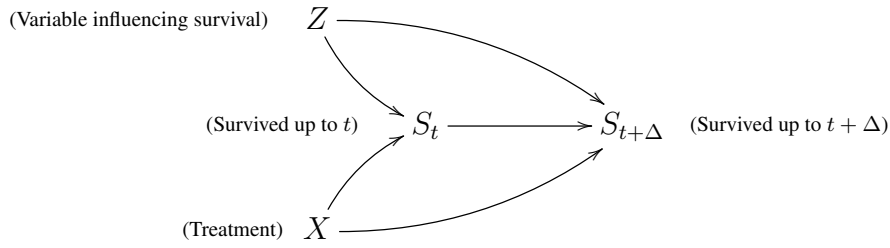


Figure 2: A directed acyclic graph of  $X$ ,  $Z$  and  $S_t$  and  $S_{t+\Delta}$

If, however, we consider the probability of surviving up to time  $t + \Delta$ , *conditional on survival up to time  $t$* , then the situation is changed. This change arises because we condition on a collider  $S_t$ , which activates the noncausal path  $X \rightarrow S_t \leftarrow Z \rightarrow S_{t+\Delta}$ . If  $Z$  is (in whole or partially) unknown, this path cannot be closed. This implies that we generally have  $X \not\perp\!\!\!\perp Z \mid T > t$ , so the compositions of the groups of treated and non-treated survivors at time  $t$  differ systematically, even if the treatment was randomly assigned at  $t = 0$ . This is a problem if we wish to assign meaning to differences of the respective hazard rates at time  $t$  since the hazards at time  $t$  are sensitive to previous survival in the two groups.

It is not automatically the case that conditioning on colliders breaks the randomization, and to clarify the conditions for this, our argument has to be supplied with a calculation as is done in Section 2.1. If the additivity condition in equation (4) is fulfilled, then we do indeed have  $X \perp\!\!\!\perp Z \mid T > t$ , so the randomization will not be broken by restricting to survivors at  $t$ .

In order to assess short-term treatment effects, not as sensitive to such systematically diverging group compositions, we imagine a hypothetical variant of our initial trial. Whenever an individual dies before  $t$ , we replace him by an identical individual that is still alive, and do not record the death that occurred before  $t$ . Then we compare the risk of deaths during  $(t, t + \Delta]$  for the treated vs. the non-treated individuals. This would provide a comparison of short-term treatment effects at time  $t$ , not sensitive to selection effects due to previous deaths, since the groups are identical to those that were formed by randomization at baseline.

In terms of causal inference, this corresponds to what is known as the controlled direct effect of treatment (Pearl, 2009), and equals:

$$\begin{aligned} \theta_x(t) &= \lim_{\Delta \rightarrow 0} \Delta^{-1} \cdot P(S_{t+\Delta} = 0 \mid \text{do}(X = x, S_t = 1)) \\ &= \lim_{\Delta \rightarrow 0} \Delta^{-1} \cdot \int_z P(S_{t+\Delta} = 0 \mid X = x, S_t = 1, z) P(dz), \end{aligned}$$

where the last expression follows from Pearl (2009, equation (3.19)). The open path  $S_t \leftarrow Z \rightarrow S_{t+\Delta}$  means that we generally have that  $P(S_{t+\Delta} = 0 \mid \text{do}(x, S_t = 1)) \neq P(S_{t+\Delta} = 0 \mid X = x, S_t = 1)$ , so our comparison can not be carried out by a straight-forward regression analysis among the survivors at time  $t$ . This has also been pointed out by Hernán et al. (2004) and Hernán and Robins (2015).

One could also take another point of view where instead of considering the controlled direct effect we consider the causal effect of treatment on short term survival conditionally on  $Z$ . By conditioning we achieve that the causal effect of treatment shall correspond to the effect it has on an individual with a given value  $Z = z$ . Let the hazard rates for an individual with  $Z = z$  be as follows under the control ( $X = 0$ ) and experimental treatment ( $X = 1$ ) conditions respectively:

$$\beta_0(t) = z \alpha(t), \quad \beta_1(t) = z r \alpha(t), \quad (5)$$

Hence, with the active treatment the individual will have  $r$  times the risk that he would have in the control group. The relationship to the previous formulation is that the treatment  $X$  has two possible options, and that for each individual the hazard is  $r$  times as large for one treatment option as for the other.

Note that the conditioning with respect to  $Z$  closes the collider path  $X \rightarrow S_t \leftarrow Z \rightarrow S_{t+\Delta}$  in Figure 2. The path  $X \rightarrow S_t \rightarrow S_{t+\Delta}$  is already closed because we condition with respect to survival at time  $t$ . Using (5) we can calculate the controlled direct effect as follows:

$$\begin{aligned}\theta_x(t) &= \int_z \lim_{\Delta \rightarrow 0} \Delta^{-1} \cdot P(S_{t+\Delta} = 0 | X = x, S_t = 1, z) P(dz) \\ &= \int_z z r^x \alpha(t) P(dz) = E(Z) r^x \alpha(t)\end{aligned}$$

Thus we get that the controlled direct effect equals  $\theta_1(t)/\theta_0(t) = r$ .

We can also calculate the other causal parameter, i.e. the one conditional on  $Z$ . To identify this, note that  $\lambda(t|do(x), z) = \lambda(t|x, z)$  where  $\lambda(t|\cdot)$  denotes the intensity of an event given various pieces of information. Hence we have from (5):

$$\frac{\lambda(t|do(X=1), Z)}{\lambda(t|do(X=0), Z)} = \frac{\lambda(t|X=1, Z)}{\lambda(t|X=0, Z)} = \frac{Z r \alpha(t)}{Z \alpha(t)} = r.$$

Hence, the causal hazard ratios defined by the controlled direct effect of treatment or by the conditional treatment effect given  $Z$  are both equal to  $r$  in this particular setting. This is not what is estimated by a Cox model in the presence of random variation in  $Z$ .

## 4 CAUSALITY AND FRAILTY

Consider the counterfactual model (5) which, while similar in form to a standard frailty model, we specify here as a basis for causal reasoning. In (5) we conceptualize the *individual* hazard rates under both treatment schemes and define the causal effect of treatment in terms of these. In practice, of course, a person will normally be assigned to one treatment, making the other assignment counterfactual. In the counterfactual framework this means that (5) holds *for any value* of  $Z$  that an individual might have. In randomized trials, however, we do not typically adjust for covariates on the presumption that randomization has distributed them (more or less) equally between the treatment groups. This corresponds to basing comparisons on the population average (marginal) survival distributions, obtained after having implicitly averaged over  $Z$ ; the effect of treatment at the individual level is not specified.

The difficulty in assigning a causal interpretation to treatment effects can be related to known phenomenon arising in frailty theory where it is well known that for certain types of frailty distributions (e.g. the compound Poisson distributions with positive probability of a zero frailty), population (marginal) hazard functions may cross over purely as an artefact of unexplained heterogeneity in the population (Flanders and Klein, 2007, Aalen et al., 2008). Hence even when a treatment is highly effective in lowering the hazard *at the individual level*, at some time the instantaneous risk at the population level becomes higher in the treatment group than the control group. This phenomenon arises in spite of the fact that for each value of the frailty variable (and hence for each individual) there is a common and constant multiplicative effect of treatment on the hazard. A test statistic reflecting the difference between treatment groups based on hazard rates (e.g. the logrank test) would therefore increase as follow-up increases to a certain point and then decrease. Although there is a mathematical connection to frailty theory, the causal reasoning we put forward here gives new insight into the concept of frailty and the phenomena that have been studied in this setting.



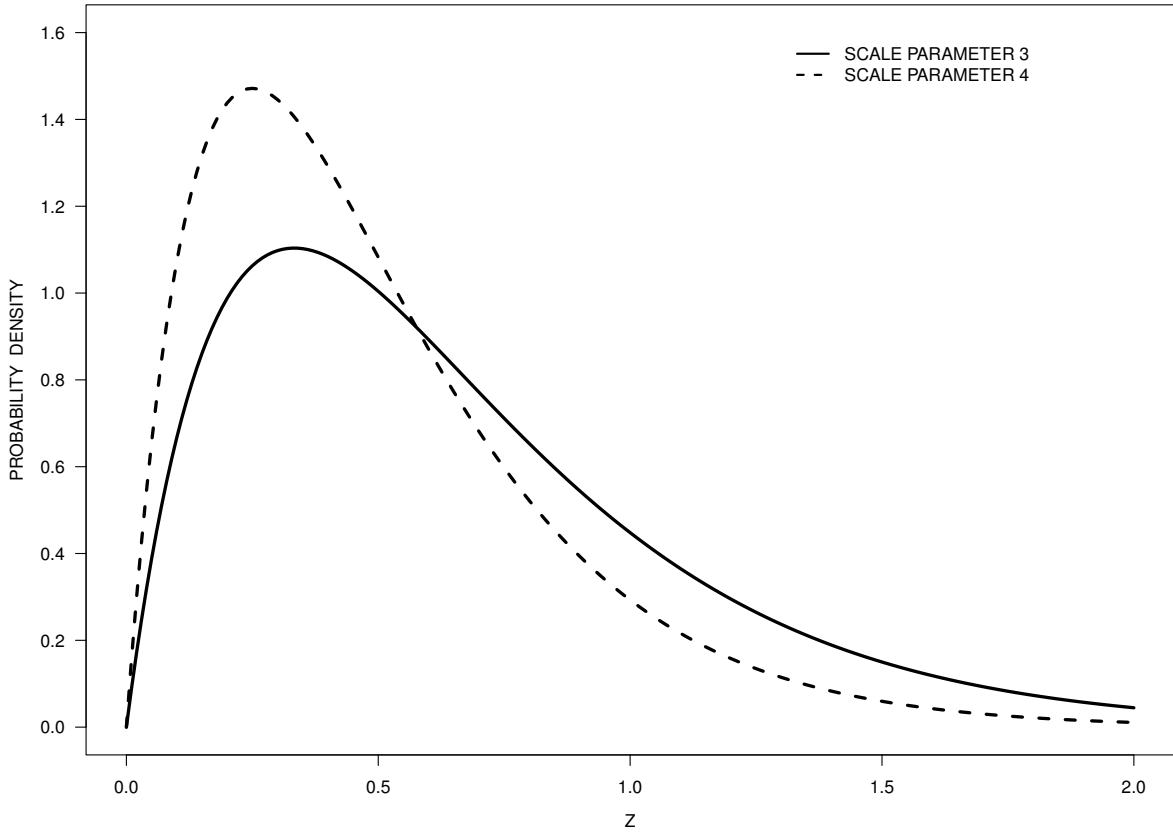


Figure 3: Two gamma distributions with shape parameter 2 and scale parameters 3 and 4 respectively, corresponding to the density of  $Z$  among survivors at time  $t$  with  $\delta = 1$ ,  $A(t) = 2$  and with a causal effect  $r = 1.5$  as specified in equation (5)

To illustrate this point further, assume that  $Z$  is gamma distributed with expectation 1 (without loss of generality) and variance  $\delta$ . Corresponding to the model in (5), the population hazard rates obtained by integrating out  $Z$  are given for the two treatment groups as follows:

$$\mu_0(t) = \frac{\alpha(t)}{1 + \delta A(t)}, \quad \mu_1(t) = \frac{r\alpha(t)}{1 + \delta r A(t)} \quad (6)$$

where  $A(t) = \int_0^t \alpha(s) ds$ . These are standard formulas, see e.g. Aalen et al. (2008). Note, that in our terminology  $\mu_0(t) = \lambda(t|X = 0)$  and  $\mu_1(t) = \lambda(t|X = 1)$ .

The ratio of the two population hazard rates is:

$$R(t) = \frac{\mu_1(t)}{\mu_0(t)} = r \cdot \left( \frac{1 + \delta A(t)}{1 + \delta r A(t)} \right). \quad (7)$$

This hazard ratio converges towards 1 in contrast to the constant causal hazard ratio  $r$  for the counterfactual model. The trend in  $R(t)$  is due to increasingly different distributions of  $Z$  for survivors in the two groups. Figure 3 contains an illustrative plot of the gamma density of  $Z$  among survivors at time  $t$  where each individual survival time has a cumulative hazard of  $A(t)$  or  $rA(t)$ ; the conditional densities have the same shape parameter  $1/\delta$  and scale parameters  $1/\delta + A(t)$  or  $1/\delta + rA(t)$  respectively.

In this setting we consider treatment effects that are not subject to selection due to frailty. To sum up previous results, we have:

$$\frac{\theta_1(t)}{\theta_0(t)} = r = \frac{\lambda(t|do(X=1), Z)}{\lambda(t|do(X=0), Z)} = \frac{\lambda(t|X=1, Z)}{\lambda(t|X=0, Z)} \neq \frac{\lambda(t|X=1)}{\lambda(t|X=0)} = r \cdot \left( \frac{1 + \delta A(t)}{1 + r\delta A(t)} \right)$$

In a randomized trial when fitting a marginal model with just the treatment as a covariate, what one estimates in a Cox model is not this causal quantity  $r$ , but a weighted average of  $R(t)$  (Lin and Wei, 1989). This will typically be closer to 1.

Can one adjust for covariates to resolve this issue? Some have advocated that baseline covariates should be controlled for in randomized trials (Hauck et al., 1998) but there is considerable discussion and debate about if, when, and how this should be carried out. The point is that the information in  $Z$  will at best only be partially known. There will with necessity be a number of dissimilarities between individuals which are unknown and even unobservable. This is the issue of frailty theory. So, in general the quantity  $r$  is not really estimable even from a randomized study.

We illustrate this by a small simulation without censoring. Assume there are two treatment groups with 1000 individuals in each group. Conditional on  $Z$  the hazard rate is  $Z$  and  $2Z$  in the two treatment groups. Assume that  $Z$  is gamma distributed with scale parameter 1 and shape parameter  $a$ . In this case the individual hazard ratio is  $r = 2$ . Using a Cox model the estimated hazard ratio is 1.04 when  $\delta = 0.1$  (extremely skewed  $Z$ ), while it is 1.36 when  $\delta = 1$  (exponential distribution for  $Z$ ). Hence, we do not get the correct individual hazard rate, but something that is strongly influenced by the general variation in risk among individuals.

Finally, in order to emphasize the importance of interventions, we examine the issue of treatment switching, discussed briefly in Aalen et al. (2008). We shall show how the effect of interventions after time 0 are misrepresented in the statistical model disregarding the frailty, and that a causal understanding is necessary. For that purpose we again use a gamma frailty model. Imagine that at some time we intervene and switch the treatment group back to the control treatment. For instance, one might observe that the hazard in two groups become very similar at some time and might wonder whether there is still a point in giving the experimental treatment. We assume that switching the treatment has an immediate effect at the individual level, meaning that the hazard for the treatment group, conditional on  $Z$ , changes from  $Zr\alpha(t)$  to  $Z\alpha(t)$  at some time  $t_1$ . Thus up to  $t_1$  the two population (marginal) hazards are given in formula (6), but after  $t_1$  the population hazard rate for group 1 is:

$$\mu_1(t) = \frac{\alpha(t)}{1 + r\delta A(t_1) + \delta(A(t) - A(t_1))}, \quad t > t_1$$

The relative population hazard for  $t > t_1$  is:

$$\begin{aligned} \frac{\mu_1(t)}{\mu_0(t)} &= \frac{1 + \delta A(t)}{1 + r\delta A(t_1) + \delta(A(t) - A(t_1))} \\ &= 1 + (1 - r) \frac{\delta A(t_1)}{1 + r\delta A(t_1) + \delta(A(t) - A(t_1))}. \end{aligned}$$

This should be compared to the relative population hazard for  $t \leq t_1$ , which is  $R(t)$ . So,

$$\frac{\mu_1(t)}{\mu_0(t)} = r \frac{1 + \delta A(t)}{1 + r\delta A(t)} = 1 - (1 - r) \frac{1}{1 + r\delta A(t)}.$$

for  $t \leq t_1$ .

Assuming  $r < 1$ , it follows that  $\mu_1(t)/\mu_0(t)$  is smaller than 1 before time  $t_1$  and larger than 1 afterwards, with a jump at  $t_1$ . Hence, the treatment group will suddenly have a higher population (marginal) hazard than the control group when treatment is discontinued, which means that the causal

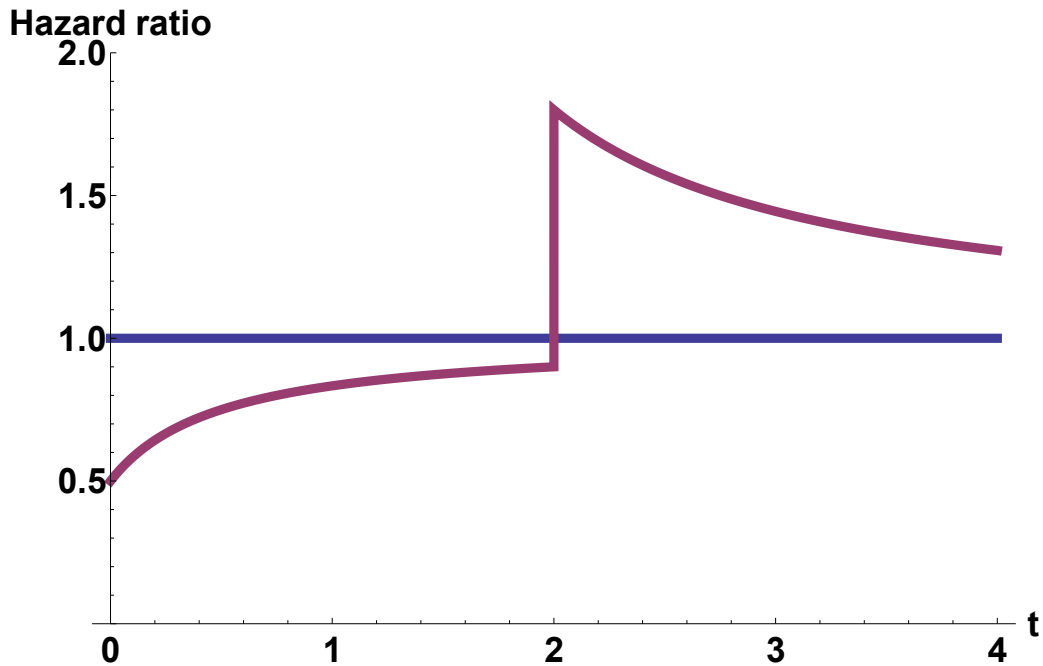


Figure 4: Hazard ratio when intervention is switched back to control at time 2.

effect of changing treatment cannot be discerned from the observed hazard rates prior to intervention. The result is illustrated in Figure 4 where  $t_1 = 2$ ,  $\alpha(t) = 1$ ,  $r = 0.5$  and  $\delta = 2$ . The treatment is switched back to control when the hazard ratio is close to 1 indicating no great difference in risk between the two groups, but the result of the intervention is a surprising and large jump.

Causality is about understanding the effect of interventions. This example shows that the population hazard rates do not necessarily give this causal insight, and that a correct causal understanding is challenging.

## 5 DISCUSSION

A central point of this note is that the hazard ratio in a Cox regression model is not the natural causal quantity to consider. This has been discussed by Greenland (1996), Hernán (2010), and Hernán et al. (2004), among others, but it does not seem widely appreciated. In light of the structural problems present with only fixed baseline variables, the current focus on the use of accelerated failure time models (Wei, 1992), models based on time-transformations (Cheng et al., 1995, Lin et al., 2014), and additive hazards models (Aalen, 1989) may be more appropriate.

In causal inference we do not presume there to be no confounding variables, but rather that they are suitably dealt with in analysis in order to mitigate their effects. Randomization achieves this goal in many settings by rendering the known and unknown confounding variables independent of the treatment indicator. With the linear model this is sufficient to ensure that the model including only the treatment indicator yields an estimator consistent for the marginal causal effect and one can likewise define the effect of interest for binary data. The Cox model features a greater structure, however, and is not collapsible (Martinussen and Vansteelandt, 2013). The issue discussed here is relevant for other methods that are based on the hazard rate. For additive hazard based models, however, the independence of  $X$  and  $Z$  in Section 2.1 is naturally preserved making it another appealing framework for causal inference like the location-scale model (Strohmaier et al., 2014).

Paradoxically it is well-known that in clinical trials one should not carry out treatment compar-

isons by conditioning on variables realized post-randomization which may be responsive to treatment since they may be on the causal pathway to the response of interest (Kalbfleisch and Prentice, 2002). Treatment comparisons based on sub-groups of individuals defined post-randomization are likewise termed improper subgroups (Yusuf et al., 1991) and are widely known to yield invalid inferences regarding treatment effects because of the benefit of randomization is lost in such comparisons. While it may be less transparent, the same process of making treatment comparisons based on subgroups of patients defined post-randomization arises in fitting the Cox model. Indeed in this setting the risk sets at a given time are defined based on survival status to that time point, a feature clearly responsive to an effective treatment. This phenomenon is particularly important in settings where interest lies in studying the time-varying effect of treatment (e.g. Durham et al., 1999) through use of flexible regression functions.

Causal inference becomes more challenging in settings involving competing risks if cause-specific analyses are of interest. In a cancer trial, for example, interest may lie in assessing the treatment effect on tumour progression through a cause-specific proportional hazards model. In patient populations where the mortality rates are appreciable, the selection effects at any given time arise from the need to be tumour-free *and* alive, so the imbalance in confounders arises from both event types. These challenges are beyond the scope of this article but warrant attention.

The issues presented here relate to comments on the meaning of the hazard rate given in Aalen et al. (2008) where it is pointed out that despite its deceptively simple definition, clear understanding of hazard rates is often elusive. These authors provide two interpretations to the hazard rate. The first is in terms of frailty theory (Chapter 6) and the second is by applying stochastic processes (Wiener processes and Lévy processes; Chapters 10 and 11). There is no doubt that the hazard rate is an important concept for explicating how the past influences the future. However, as we and others have pointed out, differences between individuals produce selection effects over time which can make it difficult to draw clear causal conclusions in this framework.

## REFERENCES

- Aalen, O., Røysland, K., Gran, J., Kouyos, R., and Lange, T. (2014). Can we believe the dags? a comment on the relationship between causal dags and mechanisms. *Statistical methods in medical research*, page 0962280213520436.
- Aalen, O. O. (1989). A linear regression model for the analysis of life times. *Statistics in Medicine*, 8(8):907–925.
- Aalen, O. O., Borgan, O., and Gjessing, H. K. (2008). *Survival and Event History Analysis: a Process Point of View*. Springer, New York, New York.
- Brown, B. M. and Wang, Y.-G. (2005). Standard errors and covariance matrices for smoothed rank estimators. *Biometrika*, 92(1):149–158.
- Cheng, S. C., Wei, L. J., and Ying, Z. (1995). Analysis of transformation models with censored data. *Biometrika*, 82(4):835–845.
- Cox, D. R. (1972). Survival models and life tables (with discussion). *Journal of the Royal Statistical Society, Series B*, 34:187–220.
- Cox, D. R. and Oakes, D. (1984). *Analysis of Survival Data*. Chapman & Hall/CRC, Boca Raton, Florida.

- Durham, L. K., Halloran, M. E., Longini, I. M., and Manatunga, A. K. (1999). Comparison of two smoothing methods for exploring waning vaccine effects. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 48(3):395–407.
- Flanders, W. D. and Klein, M. (2007). Properties of 2 counterfactual effect definitions of a point exposure. *Epidemiology*, 18(4):453–460.
- Ford, I., Norrie, J., and Ahmadi, S. (1995). Model inconsistency, illustrated by the Cox proportional hazards model. *Statistics in Medicine*, 14(8):735–746.
- Gould, A. and Lawless, J. F. (1988). Consistency and efficiency of regression coefficient estimates in location-scale models. *Biometrika*, 75(3):535–540.
- Greenland, S. (1996). Absence of confounding does not correspond to collapsibility of the rate ratio or rate difference. *Epidemiology*, 7(5):498–501.
- Hauck, W. W., Anderson, S., and Marcus, S. M. (1998). Should we adjust for covariates in nonlinear regression analyses of randomized trials? *Controlled Clinical Trials*, 19(3):249–256.
- Hernán, M. A. (2010). The hazards of hazard ratios. *Epidemiology*, 21(1):13–15.
- Hernán, M. A., Hernández-Díaz, S., and Robins, J. M. (2004). A structural approach to selection bias. *Epidemiology*, 15(5):615–625.
- Hernán, M. A. and Robins, J. M. (2015). *Causal Inference*. Chapman & Hall/CRC, Boca Raton, Florida.
- Kalbfleisch, J. D. and Prentice, R. L. (2002). *The Statistical Analysis of Failure Time Data, 2nd Edition*. John Wiley & Sons, Hoboken, New Jersey.
- Kaplan, E. L. and Meier, P. (1958). Nonparametric estimation from incomplete observation. *Journal of the American Statistical Association*, 53:457–481.
- Lawless, J. F. (2003). *Statistical Models and Methods for Lifetime Data, 2nd Edition*. John Wiley & Sons, Hoboken, New Jersey.
- Lin, D. Y. and Wei, L. J. (1989). The robust inference for the Cox proportional hazards model. *Journal of the American Statistical Association*, 84(408):1074–1078.
- Lin, H., Li, Y., Jiang, L., and Li, G. (2014). A semiparametric linear transformation model to estimate causal effects for survival data. *The Canadian Journal of Statistics*, 42(1):18–35.
- Martinussen, T. and Vansteelandt, S. (2013). On collapsibility and confounding bias in Cox and Aalen regression models. *Lifetime Data Analysis*, 19(3):279–296.
- Pearl, J. (2009). *Causality: Models, Reasoning, and Inference, 2nd Edition*. Cambridge University Press, Cambridge, UK.
- Robins, J. (1992). Estimation of the time-dependent accelerated failure time model in the presence of confounding factors. *Biometrika*, 79(2):321–334.
- Strohmaier, S., Røysland, K., Hoff, R., Borgan, Ø., Pedersen, T., and Aalen, O. O. (2014). Dynamic path analysis - a useful tool to investigate mediation processes in clinical survival trials. *Submitted*.
- Struthers, C. A. and Kalbfleisch, J. D. (1986). Misspecified proportional hazards models. *Biometrika*, 74(2):363–369.

Wei, L. J. (1992). The accelerated failure time model: a useful alternative to the Cox regression model in survival analysis. *Statistics in Medicine*, 11(14-15):1871–1879.

Yusuf, S., Wittes, J., Probstfield, J., and Tyroler, H. A. (1991). Analysis and interpretation of treatment effects in subgroups of patients in randomized clinical trials. *Journal of the American Medical Association*, 266(1):93–98.