Research Article

Statistics in Medicine

Received 7 June 2016, Accepted 11 March 2017 Published online 6 April 2017 in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.7297

Causal inference in survival analysis using pseudo-observations

Per K. Andersen,[a*†] Elisavet Syriopoulou[a,b] and Erik T. Parner[c]

Causal inference for non-censored response variables, such as binary or quantitative outcomes, is often based on either (1) direct standardization ('G-formula') or (2) inverse probability of treatment assignment weights ('propensity score'). To do causal inference in survival analysis, one needs to address right-censoring, and often, special techniques are required for that purpose.

We will show how censoring can be dealt with 'once and for all' by means of so-called pseudo-observations when doing causal inference in survival analysis. The pseudo-observations can be used as a replacement of the outcomes without censoring when applying 'standard' causal inference methods, such as (1) or (2) earlier. We study this idea for estimating the average causal effect of a binary treatment on the survival probability, the restricted mean lifetime, and the cumulative incidence in a competing risks situation.

The methods will be illustrated in a small simulation study and via a study of patients with acute myeloid leukemia who received either myeloablative or non-myeloablative conditioning before allogeneic hematopoetic cell transplantation. We will estimate the average causal effect of the conditioning regime on outcomes such as the 3-year overall survival probability and the 3-year risk of chronic graft-versus-host disease.

Keywords: survival data; causal inference; pseudo-observations; right-censoring; propensity score; G-formula

1. Introduction

A randomized study allows investigators to draw causal conclusions concerning the effect of a treatment A on an outcome Y. This is because randomization (at least in sufficiently large studies) ensures that other variables predicting Y ('risk factors', L) have the same distribution among treated ($A = 1$) and controls ($A = 0$). In other words, the treatment groups are *exchangeable* or one could say that direct comparison of the distribution of the outcome between treated and untreated provides a 'fair comparison'. In an observational study where treatment allocation is not under control of the investigator, the situation is different and a direct comparison of the distribution of Y between treated and untreated may no longer be a fair one because of the *confounding* effect arising from possibly different distributions of L in the groups $A = 0$ and $A = 1$.

Studying assumptions under which analysis of data from observational studies may still allow causal conclusions and studying methods for how to do such an analysis are active areas of research in both statistics and epidemiology. This is nicely summarized in [1]. The mathematical framework used there for formulating both assumptions and statistical models for causal inference is that of *potential outcomes*. Here, one imagines that each subject, i in a target population has two values of the outcome variable: Y_i^1 is the outcome that would be observed if the subject received treatment 1 and Y_i^0 is the outcome that would be observed if the subject received treatment 0. Because subject i belongs at most to one of the treatment groups, at least one of the potential outcomes is *counterfactual* and will never be realized. This means

[a]*Section of Biostatistics, University of Copenhagen, Ø. Farimagsgade 5, PB 2099, DK-1014, Copenhagen, Denmark*
[b]*Department of Health Sciences, College of Medicine Biological Sciences and Psychology, University of Leicester, University Road, Leicester, LE1 7RH, U.K.*
[c]*Department of Biostatistics, University of Aarhus, Bartholins Allé 2, DK-8000 Aarhus C, Denmark*
**Correspondence to: Per K. Andersen, Section of Biostatistics, University of Copenhagen, Ø. Farimagsgade 5, PB 2099, DK-1014 Copenhagen K, Denmark.*
[†]*E-mail: pka@biostat.ku.dk*

that the *individual causal effect*, typically taken to be the difference $Y_i^1 - Y_i^0$, cannot be observed and focus in causal inference is therefore often concentrated on the *average causal effect* in the total population

$$ACE = E(Y^1) - E(Y^0). \quad (1)$$

The average causal effect, *ACE* is the difference between the mean outcome over the target population if every subject were treated and the mean outcome if every subject were a control. It should be noted that other contrasts between the distributions of Y^1 and Y^0 could be studied as possible causal effects of treatment (e.g., the causal risk ratio for a binary Y, see [1, Ch. 1]) but we will, for simplicity, focus on (1) in what follows.

There are two major approaches to causal inference for a completely observed (binary or quantitative) outcome Y. One uses the *G-formula* (or 'direct standardization') and often builds on a standard regression model (the so-called 'Q-model') for the conditional expectation $E(Y \mid A, L)$ of the outcome given treatment and confounders. The other uses *inverse probability of treatment weights* (IPTW) and builds on a model for the *propensity score* $E(A \mid L)$, that is, the conditional probability of being treated given L. The latter leads to a re-weighted data set to which a *marginal structural model* is fitted. These methods have been discussed in several publications (e.g., [1–5]).

In this paper, we will focus on situations where the outcome variable Y is *incompletely observed* because of *right-censoring*. This is typically the case in survival analysis where Y is a time-to-event outcome. In order to do causal inference in such a situation, the censoring has to be addressed and this leads to a need for special methods, because methods for estimating *ACE* for binary or quantitative Y typically require that the outcome Y_i is observed for every member of the sample. If the mean parameter of interest is the survival probability $S(t_0) = E\big(I(Y > t_0)\big)$ at a given time point t_0 then both use of the G-formula where the Q-model is a Cox regression model and IPTW methods where the marginal structural model is a Cox model have been developed (e.g., [6, 7]) though other models such as additive hazard models or accelerated failure time models have also been studied (e.g., [8]). We will discuss how causal inference for parameters like $S(t_0)$ may be achieved using *pseudo-observations* (or pseudo-*values*) as the outcome in either the Q-model or in the marginal structural model fitted to the re-weighted data set. In this way, censoring is dealt with 'once and for all' and followed by standard uses of methods for a completely observed Y. This approach has the advantage that when focus is on a single time point t_0, an assumption like the proportional hazards assumption for all values t of time that is imposed when using a Cox model can be avoided.

It should be noticed that a completely different approach to causal inference, namely, *instrumental variables* (e.g., [1, Ch. 16]; [9]) has been put forward. Instrumental variable methods build on a rather different set of assumptions than those on which the approaches discussed in this paper build (Section 2). The use of pseudo-observations for causal inference in survival analysis via instrumental variables has also been discussed [10].

The structure of the paper is as follows. Section 2 gives a brief description of methods for causal inference for completely observed outcomes. Section 3 discusses causal inference for survival data and introduces pseudo-observations and how they may be used as targets. In Section 4, we present a small simulation study and a worked example from bone marrow transplantation [11], and we will compare results based on pseudo-observations with those based on previously suggested methods. Finally, Section 5 contains a brief discussion of our findings.

2. Causal inference for completely observed outcomes

Let $(Y_i^0, Y_i^1, i = 1, \ldots, n)$ be the potential outcomes in a sample of size n from the target population for whom the outcome, Y_i, the given treatment $A_i \in \{0, 1\}$ (not randomized) and confounders L_i are observed. A number of assumptions are required when using the methods to be described in the following. We assume that the treatment corresponds to a *well-defined intervention* for which a randomized trial could have been performed (see [1, Ch. 3] for a detailed discussion of 'well-defined interventions'), and to relate the observed and potential outcomes we need *consistency*, that is, $Y_i = Y_i^{A_i}$. Further, we must assume *positivity*, that is, for every value of L_i there should be a positive probability of seeing both treatments, and *no interference*, that is, Y_i^a is unaffected by setting the treatment to a^* for another subject j. Finally, an assumption of *conditional exchangeability (no unmeasured confounders)* must be imposed, that is,

$$(Y_i^0, Y_i^1) \perp\!\!\!\perp A_i \mid L_i.$$

Statistics in Medicine

The interpretation is that sufficiently many confounders are collected in L_i to provide a 'fair comparison' between treated and controls for any given value of L.

With these assumptions, the average causal effect can be estimated, because for $a = 0, 1$, a simple calculation (Section 3) gives

$$E(Y_i^a) = E_L(E(Y_i \mid L_i, A_i = a))$$

and the idea in the G-formula is then to fit a model for $E(Y_i \mid L_i, A_i)$ (the Q-model), predict the potential outcomes for subject i by $\hat{E}(Y_i \mid L_i, a), a = 0, 1$, and estimate ACE by

$$\widehat{ACE}_G = \frac{1}{n}\sum_{i=1}^{n}(\hat{E}(Y_i \mid L_i, 1) - \hat{E}(Y_i \mid L_i, 0)). \tag{2}$$

It is seen that the computation corresponds to a *direct standardization* of the predicted outcomes when treated and when being a control, where the standard distribution of the confounder L is the empirical distribution, that is, the observed distribution in the entire sample. Uncertainty of $\widehat{ACE}_G$ can be assessed by the non-parametric bootstrap, resampling with replacement from the records in the data set.

The other standard approach for causal inference is based on IPTW and first estimates the *propensity score*

$$e(L_i) = P(A_i = 1 \mid L_i),$$

that is, the conditional probability of treatment assignment given confounders. Next, a re-weighted data set is constructed replacing the outcome Y_i by $\hat{w}_i Y_i$ using the weights

$$\hat{w}_i = \frac{A_i}{\hat{e}(L_i)} + \frac{1 - A_i}{1 - \hat{e}(L_i)}.$$

By consistency and conditional exchangeability, another simple calculation (Section 3) gives that

$$E\left(\frac{1}{n}\sum_{i:A_i=1} w_i Y_i\right) = E(Y^1),$$

and similarly

$$E(Y^0) = E\left(\frac{1}{n}\sum_{i:A_i=0} w_i Y_i\right).$$

This leads to the estimator

$$\widehat{ACE}_P = \frac{1}{n}\sum_{i:A_i=1} \hat{w}_i Y_i - \frac{1}{n}\sum_{i:A_i=0} \hat{w}_i Y_i \tag{3}$$

for ACE, the idea being that the re-weighted data set is free of confounding by L and that the simple averages, therefore, possess causal interpretations. Again, uncertainty of $\widehat{ACE}_P$ can be assessed by the non-parametric bootstrap.

The two estimators (2) and (3) are identical in cases where the Q-model and the propensity score model are both *saturated*, that is, L is categorical and all interactions between its components are included in the models ([1, Ch. 2]; [12]). In general, however, the estimators are different and rely on different assumptions (in addition to those summarized in the beginning of Section 2), namely, correctness of either the Q-model or the propensity score model. *Doubly robust* estimators (relying on correctness of only one of these two models) have also been developed (e.g., [13, 14]; [1, Ch. 13]). The estimator (3) is unstable if some weights become large ($\hat{e}(L_i)$ close to 0 or 1) and *stabilized* weights

$$w_i^S = \frac{A_i \bar{A}}{\hat{e}(L_i)} + \frac{(1 - A_i)(1 - \bar{A})}{1 - \hat{e}(L_i)}$$

may be used instead, where $\bar{A}$ is the estimated *marginal* probability of treatment assignment [1, Ch. 12].

Both of the estimators (2) and (3) are estimating the treatment effect, δ_1 in the *marginal structural model*

$$E(Y^a) = \delta_0 + \delta_1 \cdot a,$$

that is, a model for the marginal distribution of the potential outcomes [1, Ch. 12-14] [2].

We have focused on *ACE*, the average causal effect in the *total population* (1), but by standardizing to other confounder distributions or using other weights, the average causal effect in sub-groups can be estimated, for example, that among the treated [1, 15, 16].

3. Causal inference for survival data

When the outcome Y is a time to an event, it is rarely completely observed for all subjects in the sample. Rather, for some subjects, only a lower limit for the event time is known – *right-censoring*. This means that *ACE* cannot be estimated directly using (2) and (3), and alternative methods are needed.

Let $Y^a, a = 0, 1$, be the potential outcomes for a survival time. We will study three different mean value parameters and the associated *ACE*. The *risk difference* at time t_0 is

$$E\big(I(Y^1 \leqslant t_0)\big) - E\big(I(Y^0 \leqslant t_0)\big),$$

the *difference in t_0-restricted mean life time* is

$$E(Y^1 \wedge t_0) - E(Y^0 \wedge t_0),$$

and with $D^a, a = 0, 1$, being the potential causes of failure in a situation with k competing causes, (i.e., D^a takes the possible values $j = 1, \ldots, k$) the *cause-j risk difference* at t_0 is

$$E\big(I(Y^1 \leqslant t_0, D^1 = j)\big) - E\big(I(Y^0 \leqslant t_0, D^0 = j)\big).$$

3.1. Classical methods

If the mean value parameter of interest is the t_0-year failure risk $1 - S(t_0) = E(I(Y \leqslant t_0))$ then (2) may be used 'indirectly' by estimating $S(t_0 \mid L, A)$, for example, via a Cox Q-model. Thus, if $\lambda_0(t)\exp(\beta^{\mathrm{T}}L + \gamma A)$ is the hazard function for that model then the estimate of $S(\cdot)$ is

$$\hat{S}(t_0 \mid L, a) = \exp\big(-\hat{\Lambda}_0(t_0)\exp(\hat{\beta}^{\mathrm{T}}L + \hat{\gamma}a)\big), \tag{4}$$

where $\Lambda_0(\cdot)$ is the cumulative baseline hazard. Along the same lines, a marginal structural Cox model

$$\delta^a(t) = \delta_0(t)\exp(\delta_1 a)$$

fitted to a propensity score re-weighted data set may lead to an estimated *ACE*

$$(1 - \exp(-\hat{\Delta}_0(t_0)\exp(\hat{\delta}_1))) - (1 - \exp(-\hat{\Delta}_0(t_0))),$$

where $\Delta_0(\cdot)$ is the cumulative baseline hazard in the marginal structural Cox model. Similar *plug-in* estimators for this *ACE* may be obtained by replacing the Cox model by other hazard models.

This kind of plug-in estimators may also be used for the *ACE* when the mean value parameter of interest is the t_0-year *restricted mean life time*

$$E(Y \wedge t_0) = \int_0^{t_0} S(t)\mathrm{d}t.$$

In a competing risks situation with causes of failure labelled $j = 1, \ldots, k$, the cumulative incidence is $F_j(t) = E\big(I(Y \leqslant t, D = j)\big)$ where D is the failure indicator. Causal inference for the *ACE* corresponding to $F_j(t_0)$ may be performed, for example, by using Cox models for all the cause-specific hazards or a Fine–Gray model for the cause j cumulative incidence as Q-models followed by (2). Cause-specific hazard models may also be fitted to a propensity score re-weighted data set, and a similar technique is available using a Fine–Gray model following the approach of [17, Ch. 5].

It is seen that the use of these methods builds on some modelling assumptions for either the Q-model or the marginal structural model. Thus, Cox or Fine–Gray models impose both some standard assumptions for a linear predictor (linearity of effects of quantitative confounders and absence/presence of certain interactions) and assumptions of proportionality of hazards or sub-distribution hazards. Additive hazard models or accelerated failure time models build on similar assumptions and, finally, the use of IPTW requires some assumptions for the propensity score model.

Statistics in Medicine

3.2. Pseudo-observations

For parameters relating to a single time point, t_0, it would be desirable to use as a Q-model or a marginal structural model a model that directly focuses on the parameter of interest, rather than having to rely on one that imposes a certain structure for all time points, such as proportional hazards. One way to obtain this is via *pseudo-observations*.

To introduce these, let f be some function of the survival time Y for which we are interested in the parameter $\theta = E(f(Y))$. Suppose that a consistent estimator $\hat{\theta}$ is available based on right-censored observations of Y for subjects $i = 1, \ldots, n$ in the sample, for example, the Kaplan–Meier estimator corresponding to $f(y) = I(y > t_0)$, or the Aalen–Johansen estimator for the cumulative incidence corresponding to $f(y) = I(y \leqslant t_0, D = j)$. The pseudo-observations for subject i is then defined as

$$\theta_i = n \cdot \hat{\theta} - (n-1) \cdot \hat{\theta}^{-i}, \tag{5}$$

where $\hat{\theta}^{-i}$ is the estimator applied to the sample of size $n-1$ obtained by eliminating subject i from the whole sample (e.g., [18, 19]). In the case of no censoring, a consistent estimator is the simple average $\hat{\theta} = (1/n)\sum_i f(Y_i)$ in which case, the pseudo-observation is simply $\theta_i = f(Y_i)$. In general, one may then think of the pseudo-observation θ_i as a replacement for the, possibly incompletely observed, random variable $f(Y_i)$. This may be a useful way forward because it has been shown [20–22] that the pseudo-observation has (approximately as $n \to \infty$, and uniformly in $i = 1, \ldots, n$) the correct conditional expectation

$$E(\theta_i \mid L_i, A_i) \approx E(f(Y_i) \mid L_i, A_i) \tag{6}$$

in situations where censoring does not depend on L, A. In cases where censoring does depend on covariates, Binder *et al.* [23] suggested alternative (inverse probability of censoring weighted) estimators for the marginal expectation θ on which computation of pseudo-observations may then be based using (5).

To make inference on the average causal effect

$$E(f(Y^1)) - E(f(Y^0))$$

we propose to use pseudo-observations θ_i for $f(Y_i)$ in (2) or (3) followed by a non-parametric bootstrap to obtain confidence limits. To show that (2) works, note that

$$E(f(Y_i^a)) = E_L(E(f(Y_i^a) \mid L_i)) = E_L(E(f(Y_i^a) \mid L_i, A_i = a))$$

(by conditional exchangeability) and, by consistency, this equals

$$E(f(Y_i^a)) = E_L(E(f(Y_i) \mid L_i, A_i = a)) \approx E_L(E(\theta_i \mid L_i, A_i = a))$$

where the last approximate equality follows from (6).

Similar arguments apply for (3), thus

$$E\left(\frac{1}{n}\sum_{i:A_i=1} w_i\theta_i\right) = \frac{1}{n}E_L\sum_i E\left(\frac{A_i\theta_i}{e(L_i)} \mid L_i\right).$$

By (6), this is (approximately)

$$\frac{1}{n}E_L\sum_i \frac{1}{e(L_i)}E\big(A_i f(Y_i) \mid L_i\big) = \frac{1}{n}E_L\sum_i \frac{1}{e(L_i)}E\big(A_i f(Y_i^1) \mid L_i\big)$$

(by consistency), and finally, (by conditional exchangeability) this is

$$\frac{1}{n}E_L\sum_i \frac{1}{e(L_i)}E\big(A_i \mid L_i\big)E\big(f(Y_i^1) \mid L_i\big) = \frac{1}{n}\sum_i E(f(Y_i^1)).$$

A similar argument applies to $E\big((1/n)\sum_{i:A_i=0} w_i\theta_i\big)$.

Compared with using the classical methods, the assumptions when using pseudo-observations are somewhat weaker. Both methods build on a correctly specified linear predictor for the Q-model or the propensity score model but the pseudo-observation method avoids assumptions of proportional hazards or proportional sub-distribution hazards. Simple uses of the method do, however, require that censoring is independent of covariates but, following [23], this extra assumption may be avoided. In the next section, we will study the impact of these assumptions in a small simulation study

4. Numerical studies

4.1. Monte Carlo simulations

We will compare the performance of the 'classical' methods for causal inference in survival analysis with methods based on pseudo-observations in a simulation study. Three scenarios (a), (b), and (c) are considered for estimating an *ACE* that is the *causal risk difference* at a fixed time point $t_0 = 1$:

(a) The Q-model is a correctly specified proportional hazards Cox model. The *ACE* is estimated either using the G-formula on the basis of (4) or on the basis of pseudo-observations. Censoring is independent. In this scenario (where we expect both methods to be unbiased), we report bias, coverage and standard deviation based on 10,000 replications.
(b) The correct Q-model is a Cox model with non-proportional hazards. The *ACE* is estimated using the G-formula on the basis of either an incorrectly specified proportional hazards Cox Q-model or on pseudo-observations. Censoring is independent. Here, we focus on a comparison of the bias for the two methods that we report on the basis of a single sample of size 1,000,000.
(c) The Q-model is a correctly specified proportional hazards Cox model. The *ACE* is estimated either using the G-formula on the basis of (4) or on the basis of pseudo-observations. Censoring is dependent on treatment, and two different sets of pseudo-observations are studied: one (incorrectly) assuming independent censoring ('Pseudo 1') and one where treatment-dependent censoring is accounted for by computing pseudo-observations on the basis of treatment-stratified Kaplan–Meier estimators ('Pseudo 2'). Again, we focus on bias that we report on the basis of a single sample of size 1,000,000.

4.1.1. Scenario (a):. The baseline hazard was constant = 0.511 for the unexposed ($A = 0$), and the hazard ratio for exposure was $HR = 1.37$. The exposure distribution was binomial with probability 0.50. Two further Gaussian variables L_1 and L_2 with a standard deviation of 0.2 were included with a hazard ratio of 1.1 per unit. This corresponds to a true *ACE* of $0.104 = 0.504 - 0.400$ at time $t_0 = 1$ (computed using 5,000,000 replications). The censoring time has a constant rate of 0.31, corresponding to an observed censoring frequency of 20.3% before time $t_0 = 1$ (computed using 5,000,000 replications). Subjects still at risk at $t_0 = 1$ were (administratively) censored at that time. Coverage, average bias in *ACE* at time 1, and standard deviation (SD) are seen in Table I. It is seen that both methods yield unbiased estimates with those based on the Cox model having a slightly smaller SD. This is what one would expect because the pseudo-observations only utilize data at time 1 whereas the Cox model (correctly) utilizes all data between times 0 and 1.

4.1.2. Scenario (b):. The baseline hazard was constant = 0.511 for the unexposed ($A = 0$) and the hazard ratio for the exposed ($A = 1$) increased linearly over time from $HR(0) = 1$ at time 0 to a hazard

Table I. Coverage, average bias ('Bias') and standard deviation ('SD') of the estimated *ACE* at time 1 computed with 10,000 replications for $n = 200, 400, 600, 800, 1000$ in scenario (a).

	Cox model			Pseudo-observations		
n	Coverage	Bias	SD	Coverage	Bias	SD
200	0.946	0.000	0.074	0.945	0.000	0.078
400	0.940	−0.001	0.053	0.945	0.000	0.054
600	0.948	−0.001	0.042	0.946	0.000	0.044
800	0.945	−0.001	0.037	0.945	0.000	0.038
1000	0.945	0.000	0.033	0.946	0.000	0.034

Statistics in Medicine

Table II. Average bias in *ACE* at time 1 ('Bias') for the Cox model and for pseudo-observations under scenario (b) for 1,000,000 observations; 'Cens. prob.' is the probability of being censored before time 1.

HR(1)	*ACE*	Cens. prob.	Bias (Cox model)	Bias (pseudo-observations)
1.0	0.000	0.000	0.001	0.001
1.0	0.000	0.075	0.000	0.000
1.0	0.000	0.143	0.001	0.001
1.0	0.000	0.315	0.000	0.000
1.0	0.000	0.515	0.000	0.000
1.0	0.000	0.732	0.002	0.003
1.5	0.072	0.000	−0.007	0.000
1.5	0.072	0.073	−0.007	0.000
1.5	0.072	0.141	−0.007	0.001
1.5	0.072	0.309	−0.012	0.000
1.5	0.072	0.509	−0.017	0.001
1.5	0.072	0.725	−0.025	0.001
2.0	0.135	0.000	−0.012	0.000
2.0	0.135	0.072	−0.013	0.001
2.0	0.135	0.139	−0.015	0.000
2.0	0.135	0.305	−0.022	−0.001
2.0	0.135	0.502	−0.030	0.000
2.0	0.135	0.718	−0.047	−0.001
5.0	0.384	0.000	−0.034	−0.002
5.0	0.384	0.067	−0.036	0.000
5.0	0.384	0.128	−0.040	−0.001
5.0	0.384	0.283	−0.052	0.000
5.0	0.384	0.470	−0.069	0.001
5.0	0.384	0.684	−0.107	0.001

ratio of $HR(1) = 1, 1.5, 2, 5$ at time 1. The exposure distribution was binomial with probability 0.50. The censoring had a constant hazard of 0, 0.1, 0.2, 0.5, 1.0, and 2.0. The true *ACE* and the censoring probabilities were computed using simulations with 5,000,000 replications. The average bias in *ACE* at time $t_0 = 1$ for the Cox model and for pseudo-observations under scenario (b) for 1,000,000 observations are reported in Table II. It is seen that estimates based on pseudo-observations are everywhere unbiased whereas those based on the Cox model have a bias that increases with the amount of mis-specification and also with the degree of censoring.

4.1.3. Scenario (c):. The baseline hazard was constant = 0.511 for the unexposed ($A = 0$) and the hazard ratio was $HR = 1, 2$ for the exposed ($A = 1$). The exposure distribution was binomial with probabilities $P(A = 1) = 0.10, 0.25, 0.50$. The censoring time had a constant hazard of 0.1 in the unexposed group and was 0.1, 0.2, 0.5, 1.0, and 2 in the exposed group. The true *ACE* and censoring probabilities are computed using simulations with 5,000,000 replications. The average bias in *ACE* for the Cox model, the simple incorrect pseudo-observations (Pseudo 1) and the stratified pseudo-observations (Pseudo 2) for scenario (c) for 1,000,000 observations are reported in Table III. It is seen that estimates based on the correctly specified Cox model (as in scenario (a)) are unbiased, while those based on the simple pseudo-observations, not accounting for treatment-dependent censoring have a bias that increases with the amount of discrepancy between the two censoring distributions. That bias, however, disappears when the treatment-stratified pseudo-observations are used.

4.2. Example: causal effect of conditioning in acute myeloid leukemia patients

Sengeløv *et al.* [11] reported results from follow-up of 207 consecutive patients treated in a single institution with *allogeneic haematopoeitic cell transplantation* ('bone marrow transplantation', HCT) for *acute myeloid leukemia* (AML). In preparation of HCT, patients were, in a non-randomized fashion, given either a *myeloablative* (MA) or *non-myeloablative* (NM) conditioning, and in this illustrative example, we will compare the outcome after HCT between the two conditioning regimes using the methods discussed in Sections 2-3. Out of the 207 patients, 122 were transplanted in first complete remission and we will restrict attention to the 116 patients transplanted in first complete remission and for whom information

Table III. Average bias in *ACE* at time 1 ('Bias') for the Cox model, simple (Pseudo 1) and stratified (Pseudo 2) pseudo-observations under scenario (c) for 1,000,000 observations; 'Cens. prob.' is the probability of being censored before time 1.

			Cens. prob.		Bias		
$P(A=1)$	*HR*	*ACE*	$A=0$	$A=1$	Cox model	Pseudo 1	Pseudo 2
0.10	1	0.00	0.07	0.07	−0.003	−0.006	−0.006
0.10	1	0.00	0.07	0.14	0.002	0.002	0.002
0.10	1	0.00	0.07	0.31	0.000	−0.001	−0.001
0.10	1	0.00	0.08	0.52	0.002	0.001	0.001
0.10	1	0.00	0.07	0.73	−0.001	−0.005	−0.002
0.10	2	0.24	0.08	0.06	−0.001	−0.002	−0.002
0.10	2	0.24	0.07	0.12	0.000	0.008	−0.001
0.10	2	0.24	0.07	0.26	−0.001	0.042	0.000
0.10	2	0.24	0.07	0.43	0.001	0.103	0.002
0.10	2	0.24	0.08	0.63	0.002	0.206	0.002
0.25	1	0.00	0.07	0.07	0.001	0.000	0.000
0.25	1	0.00	0.07	0.14	−0.001	−0.002	−0.002
0.25	1	0.00	0.07	0.31	−0.002	−0.002	−0.002
0.25	1	0.00	0.07	0.52	0.000	0.000	0.000
0.25	1	0.00	0.08	0.73	0.002	0.003	0.002
0.25	2	0.24	0.07	0.06	0.001	0.001	0.001
0.25	2	0.24	0.08	0.12	−0.001	0.005	−0.001
0.25	2	0.24	0.07	0.26	−0.000	0.023	0.001
0.25	2	0.24	0.08	0.43	0.002	0.041	0.001
0.25	2	0.24	0.07	0.63	0.000	0.046	−0.001
0.50	1	0.00	0.08	0.08	0.000	0.000	0.000
0.50	1	0.00	0.07	0.14	−0.001	0.001	0.001
0.50	1	0.00	0.07	0.31	0.001	−0.001	−0.001
0.50	1	0.00	0.07	0.52	0.001	0.000	0.000
0.50	1	0.00	0.07	0.73	0.001	0.000	−0.001
0.50	2	0.24	0.07	0.06	0.000	0.000	0.000
0.50	2	0.24	0.07	0.12	0.000	−0.001	0.000
0.50	2	0.24	0.07	0.26	−0.001	−0.006	−0.002
0.50	2	0.24	0.07	0.43	−0.001	−0.020	−0.003
0.50	2	0.24	0.07	0.63	−0.003	−0.056	−0.004

on the relevant potential confounders was available (six patients lacking information on the Karnofsky prognostic score were excluded). It is important to realize that our analyses are meant as being illustrative of the methods because we have not taken age of the patients into account. This is due to the fact that older patients tended to be treated with NM conditioning, thereby violating the *positivity* assumption discussed earlier. The outcomes considered by Sengeløv *et al.* [11] were: overall survival time and relapse-free survival time, the latter corresponding to no occurrence of the two competing end-points: relapse and non-relapse mortality, which were also considered as separate outcomes by Sengeløv *et al.* [11] in a competing risks analysis. Finally, occurrence of acute (aGvHD) and chronic (cGvHD) graft-versus-host disease was studied taking the competing events of relapse and non-relapse mortality into account.

For illustration, we will concentrate on the overall survival time and on the occurrence of cGvHD, and Figures 1 and 2 show, respectively, the Kaplan–Meier and Aalen–Johansen estimates in the MA and NM preconditioning groups. We will concentrate on the status 3 years after HCT where, according to the Kaplan–Meier estimates, the *risk difference* corresponding to overall survival (MA – NM) is (1 to 0.68)−(1 to 0.66)=−0.02 with a 95% confidence interval from −0.16 to 0.20. Similarly, the difference between the estimated cumulative incidences of cGvHD (MA – NM) is 0.53 − 0.41 = 0.12 (95% c.i. from −0.07 to 0.31). However, owing to the non-randomized nature of conditioning treatment allocation, these estimates may be confounded, and in what follows, we will adjust for this confounding and estimate average causal effects using the methods discussed in Sections 2-3.

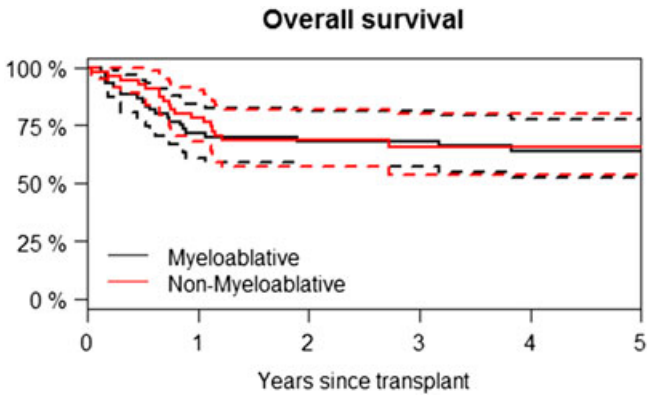

Figure 1. Kaplan–Meier estimates for overall survival by preconditioning regime, myeloablative (MA) versus non-myeloablative (NM), with 95% confidence limits (dashed). [Colour figure can be viewed at wileyonlinelibrary.com]

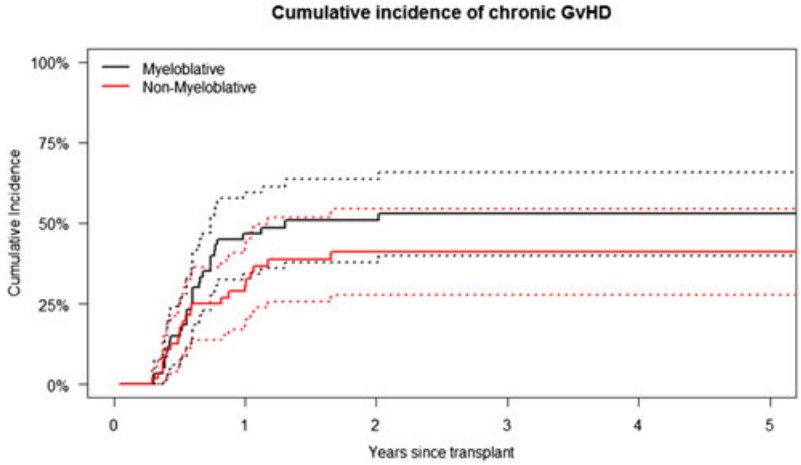

Figure 2. Aalen–Johansen estimates for the cumulative incidence of chronic graft-versus-host disease (cGvHD) by preconditioning regime, myeloablative (MA) vs. non-myeloablative (NM), with 95% confidence limits (dashed). [Colour figure can be viewed at wileyonlinelibrary.com]

Table IV. Pre-treatment characteristics (%) in NM and MA conditioning groups and estimated coefficients (log(*OR*)) in a logistic regression propensity score model (with standard error, SD).

Characteristic	NM ($n = 56$)	MA ($n = 60$)	log(*OR*)	SD
High cytogeneic risk	25	33	0.404	0.421
Male donor	54	63	0.465	0.395
Karnofsky score < 90	14	10	−0.383	0.608
No prior myelodysplasia	77	82	0.294	0.500
Sibling donor	68	67	−0.083	0.439
Same sex donor	57	55	−0.163	0.394

Table IV shows the distribution of pre-treatment characteristics in the NM and MA groups, and it is seen that more MA patients had a high cytogeneic risk, a high Karnofsky score and no prior myelodysplasia whereas other factors had more similar distributions in the two treatment groups. Table IV also shows the estimated log(odds ratios) resulting from fitting a logistic propensity score model for the probability of receiving an MA conditioning.

4.2.1. Overall survival time. To estimate the average causal effect on the three-year overall survival probability, we first fitted a Cox model including conditioning treatment and the variables in Table IV

Table V. Association between pre-treatment characteristics plus MA vs. NM conditioning and overall survival time.

	Cox model		Model for pseudo-observations	
Characteristic	*HR*	*SD*(log(*HR*))	$\hat{\beta}$	*SD*($\hat{\beta}$)
MA vs. NM	0.951	0.318	−0.025	0.090
High cytogeneic risk	1.293	0.339	−0.129	0.099
Male donor	1.244	0.334	−0.002	0.094
Karnofsky score < 90	2.192	0.421	0.252	0.143
No prior myelodysplasia	1.319	0.419	0.064	0.119
Sibling donor	0.618	0.347	−0.105	0.103
Same sex donor	1.089	0.336	0.096	0.093
ACE (MA vs. NM)	−0.013 (−0.181, 0.158)		−0.025 (−0.205, 0.150)	

Left: hazard ratios (*HR*) using a Cox model, right: risk differences ($\hat{\beta}$) using a linear model with pseudo-observations at 3 years as responses (with standard error, SD). The last line shows the estimated average causal effect (MA vs. NM) using the G-formula (2) with 95% confidence limits obtained using the bootstrap.

as covariates and used that as the Q-model to predict the 3-year overall survival probabilities for each patient under both of the conditioning regimes using (4). Second, we computed pseudo-observations at 3 years for each patient on the basis of (5), where $\hat{\theta}$ is the value at 3 years of the overall Kaplan–Meier estimator, that is, for the full sample of 116 patients. These pseudo-observations were subsequently used as response in a Q-model relating the mean response linearly to conditioning treatment and covariates. Results from these two Q-models are shown in Table V.

From either Q-model, the average causal effect on the 3-year *risk* of death was estimated using the G-formula (2), and confidence limits were obtained using the percentile method on the basis of 2000 bootstrap replications. From Table V, it is seen that the two estimates are quite similar and not very different from the unadjusted estimate on the basis of the Kaplan–Meier estimators. The similarity with the unadjusted value is owing to the low degree of confounding also suggested by the results in Table IV where no covariate seems to be a strong predictor for preconditioning treatment.

For comparison, we also fitted a marginal structural model to the data set where pseudo-observations were weighted by the inverse probabilities of conditioning treatment allocation with probabilities estimated from the propensity score model in Table IV. Using (3), this resulted in an average causal effect of −0.028 with a 95% confidence interval based on 2000 bootstrap samples from −0.197 to 0.157. The corresponding estimate obtained by fitting a marginal structural Cox model to the re-weighted data set was −0.018 with 95% confidence obtained using the bootstrap ranging from −0.166 to 0.182. The estimates are close to those obtained using the G-formula.

4.2.2. Occurrence of chronic graft-versus-host disease. A similar set of analyses was carried out in order to estimate the average causal effect of preconditioning treatment on the 3-year risk of chronic graft-versus-host disease. Thus, pseudo-observations were computed on the basis of (5), where $\hat{\theta}$ is the value at 3 years of the overall Aalen–Johansen estimator of the cumulative incidence of cGvHD taking the competing risks of relapse and non-relapse mortality into account. The *ACE* was then estimated using, on the one hand, the G-formula and a marginal structural model fitted to an inverse probability of treatment weighted data set both using pseudo-observations and, on the other hand, the G-formula where the Q-model is a Fine–Gray regression model used for predicting the cumulative incidence of cGvHD for each patient at 3 years under either preconditioning treatment.

Table VI shows the estimated coefficients from the two Q-models used, that is, sub-distribution hazard ratios (*SHR*) for the Fine–Gray model and risk differences ($\hat{\beta}$) for the linear model for the pseudo-observations at 3 years.

The estimated risk differences are similar with confidence intervals of similar width and both slightly larger than the unadjusted estimated based on Figure 1. For comparison, the estimate based on a marginal structural linear model for the pseudo-observations is 0.157 with a 95% confidence interval from −0.074 to 0.307.

4.2.3. Alternative models. The Q-models used in the example were all very simple and included, for example, no interactions between the covariates. To investigate the robustness, we compared with results

Table VI. Association between pre-treatment characteristics plus MA vs. NM conditioning and risk of cGvHD.

Characteristic	Fine-Gray model		Model for pseudo-observations	
	SHR	*SD*(log(*SHR*))	$\hat{\beta}$	*SD*($\hat{\beta}$)
MA vs. NM	1.818	0.301	0.505	0.152
High cytogeneic risk	0.523	0.375	−0.166	0.101
Male donor	0.549	0.268	−0.192	0.096
Karnofsky score < 90	1.006	0.431	−0.012	0.146
No prior myelodysplasia	0.561	0.410	−0.175	0.121
Sibling donor	1.853	0.407	0.158	0.106
Same sex donor	0.606	0.303	−0.187	0.095
ACE (MA vs. NM)	0.177 (−0.005, 0.339)		0.157 (−0.016, 0.334)	

Left: sub-distribution hazard ratios (SHR) using a Fine–Gray model. Right: risk differences ($\hat{\beta}$) using a linear model with pseudo-observations as responses (with standard error, SD). The last line shows the estimated average causal effect (MA vs. NM) using the G-formula (2) with 95% confidence limits obtained using the bootstrap.

obtained by introducing an interaction between treatment and Karnofsky score in the Q-models. For 3-year overall survival, the estimated *ACE* changed from −0.013 for the Cox Q-model and −0.025 for pseudo-observations (Table V) to, respectively, −0.019 and −0.026. For the 3-year risk of cGvHD, the estimated *ACE* changed from 0.177 for the Fine–Gray Q-model and 0.157 for pseudo-observations (Table VI) to, respectively, 0.175 and 0.157. Thus, relaxing the modelling assumptions for the Q-models in this way had minimal impact on the results.

5. Discussion

We have proposed to use pseudo-observations for the purpose of doing causal inference in survival analysis. Because pseudo-observations provide a general tool for analyzing mean value parameters in the presence of right-censoring, the method lends itself to estimating an average causal effect in survival analysis. We have illustrated the use of the method for estimating the causal risk difference at a fixed point in time (possibly in the presence of competing risks) but it applies equally easily to parameters like the restricted mean life time [24] or the number of years spent in a given state in a multi-state model [25, 26]. Relying on the simple Kaplan–Meier or Aalen–Johansen estimators, a drawback is that censoring is not allowed to depend on covariates. However, Binder *et al*. [23] showed how alternative inverse probability of censoring weighted estimators may form the basis for computing the pseudo-observations in such a situation. Thus, for a correctly specified model for the distribution of the censoring time C:

$$G(t \mid L, A) = P(C > t \mid L, A),$$

the Kaplan–Meier estimator is replaced by the consistent estimator

$$\hat{S}_C(t) = 1 - \frac{1}{n} \sum_i \frac{N_i(t)}{\hat{G}(Y_i - \mid L_i, A_i)}$$

of the marginal distribution $S(t) = P(Y > t)$, where $N_i(t)$ is the counting process $I(Y_i \leqslant t, D_i = 1)$ and $\hat{G}(\cdot)$ the estimated censoring distribution. In our simulation study, we showed how this works in a simple setting where the censoring was allowed to be treatment-dependent.

A nice feature of the pseudo-observation approach is that, once censoring is dealt with in the computation of pseudo-observations, causal inference proceeds in exactly the same way as for completely observed outcomes using either the G-formula (2) or the IPTW estimator (3).

For simplicity, we have focused on the situation where A is a binary 'point treatment' given at 'time 0' and, as a consequence, *time-dependent confounding* ([3]) is not an issue. However, when the aim of a study is to estimate causal effects of a time-varying treatment ($\bar{A}_T = (A_0, A_1, \ldots, A_T)$) on an ultimate outcome $Y(T + 1)$, for example, the survival status $I(Y \geqslant T + 1)$ at time $T + 1$ and when adjustment for a time-dependent confounder $\bar{L}_T = (L_0, L_1, \ldots, L_T)$ is needed, both the G-formula and IPTW (as discussed by [3]) may be used by replacing the possibly incompletely observed outcome variable by its pseudo-observation.

In our example, the impact of using pseudo-observations for inference was minor. This is probably because of the limited degree of confounding in the example where the risk factor distributions in the two conditioning regimes were not very different (Table IV). A further drawback with the example is that one variable, age, had distributions that differed *too much* between the treatment groups, thereby violating the positivity assumption. However, in a sensitivity analysis restricting attention to the rather small group of patients within the common age range for the two treatments, results did not differ considerably from those reported earlier, albeit with greater uncertainty. The example did, however, show the feasibility of the proposed method and the ease with which causal inference may be achieved for censored data.

Compared with the proportionality assumptions needed when using a Cox model or a Fine–Gray model as a Q-model or as a marginal structural model for inference at a fixed time point, the pseudo-observation approach avoids such assumptions. Indeed, our simulations showed that a mis-specification of a Cox Q-model resulted in biased *ACE* estimates, a bias that was avoided when using pseudo-observations. It should be kept in mind, however, that all methods rely on a correct specification of the linear predictor in either the Q-model or the model for the propensity score.

Software for calculating pseudo-observations are available in both `SAS` and `R` [27] and in `Stata` [28, 29].

References

1. Hernan MA, Robins J. *Causal Inference*. CRC/Chapman and Hall: Boca Raton, 2016. (In press).
2. Robins JM, Hernan MA, Brumback B. Marginal structural models and causal inference in epidemiology. *Epidemiology* 2000; **11**(5):550–560.
3. Daniel RM, Cousens SN, De Stavola BL, Kenward MG, Sterne JA. Methods for dealing with time-dependent confounding. *Statistics in Medicine* 2013; **32**(9):1584–1618.
4. Austin PC. An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behavioral Research* 2011; **46**(3):399–424.
5. Snowden JM, Rose S, Mortimer KM. Implementation of G-computation on a simulated data set: demonstration of a causal inference technique. *American Journal of Epidemiology* 2011; **173**(7):731–738.
6. Austin PC. Absolute risk reductions and numbers needed to treat can be obtained from adjusted survival models for time-to-event outcomes. *Journal of Clinical Epidemiology* 2010; **63**(1):46–55.
7. Austin PC. The use of propensity score methods with survival or time-to-event outcomes: reporting measures of effect similar to those used in randomized experiments. *Statistics in Medicine* 2014; **33**(7):1242–1258.
8. Martinussen T, Vansteelandt S. On collapsibility and confounding bias in Cox and Aalen regression models. *Lifetime Data Analysis* 2013; **19**(3):279–296.
9. Angrist JD, Imbens GW, Rubin DB. Identification of causal effects using instrumental variables. *Journal of the American Statistical Association* 1996; **91**:444–455.
10. Kjaersgaard MI, Parner ET. Instrumental variable method for time-to-event data using a pseudo-observation approach. *Biometrics* 2016; **72**(2):463–72.
11. Sengeløv H, Gerds T, Brændstrup P, Kornblit B, Mortensen B, Petersen S, Vindeløv L. Long-term survival after allogeneic haematopoietic cell transplantation for AML in remission: single-centre results after TBI-based myeloablative and non-myeloablative conditioning. *Bone Marrow Transplantation* 2013; **48**(9):1185–1191.
12. Hernan MA, Robins JM. Estimating causal effects from epidemiological data. *Journal of Epidemiology and Community Health* 2006; **60**(7):578–586.
13. Bang H, Robins JM. Doubly robust estimation in missing data and causal inference models. *Biometrics* 2005; **61**(4): 962–973. (Correction: (2008). **64**, 650.)
14. Funk MJ, Westreich D, Wiesen C, Sturmer T, Brookhart MA, Davidian M. Doubly robust estimation of causal effects. *American Journal of Epidemiology* 2011; **173**(7):761–767.
15. Kurth T, Walker AM, Glynn RJ, Chan KA, Gaziano JM, Berger K, Robins JM. Results of multivariable logistic regression, propensity matching, propensity adjustment, and propensity-based weighting under conditions of nonuniform effect. *American Journal of Epidemiology* 2006; **163**(3):262–270.
16. Sato T, Matsuyama Y. Marginal structural models as a tool for standardization. *Epidemiology* 2003; **14**(6):680–686.
17. Geskus R. *Data Analysis with Competing Risks and Intermediate States*. CRC/Chapman and Hall: Boca Raton, 2016.
18. Andersen PK, Klein JP, Rosthøj S. Generalized linear models for correlated pseudo-observations, with applications to multi-state models. *Biometrika* 2003; **90**:15–27.
19. Andersen PK, Perme MP. Pseudo-observations in survival analysis. *Statistical Methods in Medical Research* 2010; **19**(1):71–99.
20. Graw F, Gerds TA, Schumacher M. On pseudo-values for regression analysis in competing risks models. *Lifetime Data Analysis* 2009; **15**(2):241–255.
21. Jacobsen M, Martinussen T. A note on the large sample properties of estimators based on generalized linear models for correlated pseudo-observations. *Scandinavian Journal of Statistics* 2016; **43**(3):845–862.
22. Overgaard M, Parner ET, Pedersen J. Asymptotic theory of generalized estimating equations based on jack-knife pseudo-observations. *Annals of Statistics* (in press).
23. Binder N, Gerds TA, Andersen PK. Pseudo-observations for competing risks with covariate dependent censoring. *Lifetime Data Analysis* 2014; **20**(2):303–315.

24. Andersen PK, Hansen MG, Klein JP. Regression analysis of restricted mean survival time based on pseudo-observations. *Lifetime Data Analysis* 2004; **10**(4):335–350.
25. Grand MK, Putter H. Regression models for expected length of stay. *Statistics in Medicine* 2016; **35**(7):1178–1192.
26. Andersen PK. Decomposition of number of years lost according to causes of death. *Statistics in Medicine* 2013; **32**(30):5278–5285.
27. Klein JP, Gerster M, Andersen PK, Tarima S, Pohar Perme M. SAS and R functions to compute pseudo-values for censored data regression. *Computer Methods and Programs in Biomedicine* 2008; **89**:289–300.
28. Parner ET, Andersen PK. Regression analysis of censored data using pseudo observations. *The Stata Journal* 2010; **10**: 408–422.
29. Overgaard M, Andersen PK, Parner ET. Regression analysis of censored data using pseudo-observations: An update. *The Stata Journal* 2015; **15**:809–821.