Causal inference in survival analysis using pseudo-observations

Per Kragh Andersen
Elisavet Syriopoulou
Erik Parner

Research Report 16/03

Department of Biostatistics
University of Copenhagen
Causal inference in survival analysis using pseudo-observations

Per Kragh Andersen¹, Elisavet Syriopoulou¹², Erik Parner³

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

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. We will show how ‘standard’ causal inference techniques, such as (1) or (2) above, may be applied to the pseudo-observations. We consider the idea on right-censored survival data to estimate the average causal effect on the survival probability and restricted mean. The same idea applies to competing risks.

The methods will be illustrated via a study of patients with acute myeloid leukemia who received either myeloablative or non-myeloablative conditioning before allogeneic hematopoietic 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 Department of Biostatistics, University of Copenhagen, Ø. Farimagsgade 5, PB 2099, DK-1014 Copenhagen K, Denmark

² Department of Health Sciences, University of Leicester, Adrian Building, University Road, Leicester, LE1 7RH, U.K.

³ Department of Biostatistics, University of Aarhus, Bartholins Allé 2, DK-8000 Aarhus C, Denmark.
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 by Hernan and Robins (2016). 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: $Y^1_i$ is the outcome that would be observed if the subject, by randomization, were given the treatment ($A_i = 1$) and $Y^0_i$ is the outcome that would be observed if the subject, by randomization, were allocated to the control group ($A_i = 0$). Since 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 that the individual causal effect, typically taken to be the difference $Y^1_i - Y^0_i$, cannot be observed and focus in causal inference is therefore often concentrated on the average causal effect in the total population

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

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 Hernan and Robins, 2016, 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 $G$-estimation (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., Hernan and Robins, 2016; Robins et al., 2000; Daniel et al., 2013; Austin, 2011; Snowden et al.,
2011) and so have other techniques for causal inference, such as instrumental variables (e.g., Hernan and Robins, 2016, Ch. 16; Angrist et al., 1996).

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, since techniques 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 \( S(t_0) = E(I(Y > t_0)) \) at a given time point \( t_0 \) then both G-estimation methods 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., Austin, 2010; 2014) though other models such as additive hazard models or accelerated failure time models have also been studied (e.g., Martinussen and Vansteelandt, 2013). We will discuss how causal inference for parameters like \( S(t_0) \) may be achieved using *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 techniques for 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 which is imposed when using a Cox model can be avoided.

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-values and how they may be used as targets. In Section 4 we present a worked example from bone marrow transplantation (Sengeløv et al., 2013) and we will compare results based on pseudo-values with those based on previously suggested methods. The final Section 5 contains a brief discussion of our findings.

### 2 Causal inference for completely observed outcomes

Let \( (Y_i^0, Y_i^1, i = 1, ..., n) \) 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. We assume that the treatment corresponds to a *well-defined intervention* for which a randomized trial could have been performed (see Hernan and Robins, 2016, Ch. 3, for a detailed discussion of ‘well-defined interventions’). To relate the observed and potential outcomes we need to assume *consistency*, i.e. \( Y_i = Y_i^{A_i} \) (the outcome for subject \( i \) when observed to have been given treatment \( A_i \) is the same as would have been observed, had \( i \) been randomized to that treatment). Finally an assumption of *exchangeability* (no unmeasured confounders) must be
imposed, i.e.

$$(Y^0_i, Y^1_i) \perp A_i \mid L_i.$$  

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 $L_i$.

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

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

and the idea in G-estimation 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

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

For these averages (estimating the expectation over $L$ using the empirical con-
founder distribution) to estimate $E(Y^a)$ consistently a final assumption of positivity is needed, i.e. for every value of $L_i$ there should be a positive probability of seeing both treatments. It is seen that the computation corresponds to di-
rect standardization of the predicted outcomes when treated and when being a control to the entire sample and its observed distribution of $L$.

Uncertainty of $\hat{ACE}_G$ 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

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

i.e., 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

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

where, again, positivity $0 < e(L_i) < 1$ is needed. By consistency and exchange-
ability another simple calculation (see Section 3) gives that

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

and similarly

$$E(Y^0) = E(\frac{1}{n} \sum_{i: A_i = 0} w_i Y_i).$$
This leads to the estimator
\[
\hat{ACE}_P = \frac{1}{n} \sum_{i: A_i = 1} \hat{w}_i Y_i - \frac{1}{n} \sum_{i: A_i = 0} \hat{w}_i Y_i
\]
for \(ACE_P\), 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 \(ACE_P\) 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, i.e. \(L\) is categorical and all interactions between its components are included in the models (Hernan and Robins, 2016, Ch. 2; 2006). In general, however, the estimators are different and rely on different assumptions, 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., Bang and Robins, 2005; Hernan and Robins, 2016, Ch. 13; Funk et al., 2011). The estimator (3) is unstable if some weights become large (\(\hat{e}(L_i)\) close to 0 or 1) and stabilized weights
\[
\hat{w}_i^S = \frac{A_i \hat{A}}{\hat{e}(L_i)} + \frac{(1 - A_i)(1 - \hat{A})}{1 - \hat{e}(L_i)}
\]
may be used instead, where \(\hat{A}\) is the estimated marginal probability of treatment assignment (Hernan and Robins, 2016, Ch. 12).

Both of the estimators (2) and (3) are estimating the treatment effect, \(\delta_1\) in the marginal structural model
\[
E(Y^a) = \delta_0 + \delta_1 \cdot a,
\]
i.e., a model for the marginal distribution of the potential outcomes (Hernan and Robins, 2016, Ch. 12-14; Robins et al., 2000).

We have focused on \(ACE\), the average causal effect in the total population (1) but standardizing to other confounder distributions or using other weights, the average causal effect in sub-groups can be estimated, e.g. that among the treated (Hernan and Robins, 2016; Kurth et al., 2005; Sato and Matsuyama, 2003).

3 Causal inference for survival data; pseudo-observations

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.

If the mean value parameter of interest is the \(t_0\)-year survival probability \(S(t_0) = E(I(Y > t_0))\) then (2) may be used ‘indirectly’ by estimating \(S(t_0 \mid L, A)\), e.g. via a Cox Q-model, \(\lambda_0(t) \exp(\beta L + \gamma A)\) for the hazard function as
\[
\hat{S}(t_0 \mid L, a) = \exp(-\hat{\Lambda}_0(t_0) \exp(\beta^T L + \gamma a)).
\]
Along the same lines, a marginal structural Cox model
\[ \delta \alpha(t) = \delta_0(t) \exp(\delta_1 \alpha) \]
fitted to a propensity score re-weighted data set may lead to an estimated ACE
\[ \exp(-\tilde{\Delta}_0(t_0)) - \exp(-\tilde{\Delta}_0(t_0) \exp(\hat{\delta}_1)). \]

Similar plug-in estimators for ACE may be applied in other survival data situations. Thus, the Cox model may be replaced by other hazard models when estimating causal effects on the \( t_0 \)-year survival probability or on a parameter like the \( t_0 \)-year restricted mean lifetime \( \int_{t_0}^0 S(t)dt \).

In a competing risks situation with causes of failure labelled \( j = 1, \ldots, k \), the cumulative incidence is \( F_j(t) = E(I(Y \leq t, D = j)) \) (with \( D \), the failure indicator) and causal inference for \( F_j(t_0) \) may be performed, e.g. 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 Geskus (2016, Ch. 5).

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. 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, e.g. the Kaplan-Meier estimator corresponding to \( f(y) = I(y > t_0) \), or the Aalen-Johansen estimator for the cumulative incidence corresponding to \( f(y) = I(y \leq t_0, D = j) \). The pseudo-value for subject \( i \) is then defined as
\[ \theta_i = n \cdot \hat{\theta} - (n - 1) \cdot \hat{\theta}^{-i}, \]
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., Andersen et al., 2003; Andersen and Pohar Perme, 2010). In the case of no censoring a consistent estimator is \( \hat{\theta} = (1/n) \sum_i f(Y_i) \) in which case the pseudo-value is simply \( \theta_i = f(Y_i) \). One may then think of the pseudo-value \( \theta_i \) 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 (Graw et al., 2009; Jacobsen and Martinussen, 2016) that the pseudo-value has (approximately as \( n \to \infty \), and uniformly in \( i = 1, \ldots, n \)) the correct conditional expectation
\[ E(\theta_i \mid L_i, A_i) = E(f(Y_i) \mid L_i, A_i) + o_P(1) \]
in situations where censoring does not depend on \( L, A \). In cases where censoring does depend on covariates, Binder et al. (2014) suggested alternative (inverse
probability of censoring weighted) estimators for the marginal expectation \( \theta \) on which computation of pseudo-values may then be based using (4).

To make inference on the average causal effect

\[
E(f(Y^1)) - E(f(Y^0))
\]
we propose to use pseudo-values \( \theta_i \) 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 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 (5).

Similar arguments apply for (3), thus

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

By (5) this is (approximately)

\[
\frac{1}{n} E_L \sum_i \frac{1}{e(L_i)} E(A_i f(Y_i) \mid L_i) = \frac{1}{n} E_L \sum_i \frac{1}{e(L_i)} E(A_i f(Y_i^1) \mid L_i)
\]
(by consistency), and finally (by exchangeability) this is

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

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

4 Example: causal effect of conditioning in AML patients

Sengeløv et al. (2013) reported results from follow-up of 207 consecutive patients treated in a single institution with allogeneic haematopoetic 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 the previous section.

We will restrict attention to the 116 patients transplanted in first complete remission and for whom information on the relevant potential confounders was available. Thus, six patients without information on the Karnofsky prognostic score were excluded. It is important to realize that our analyses are meant as
being illustrative of the methods since we have not taken age of the patients into account. This is due the fact that older patients tended to be treated with NM conditioning, thereby violating the positivity assumption discussed above. The outcomes considered by Sengeløv et al. (2013) were: overall survival and relapse-free survival, 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. (2013) in a competing risks analysis. Finally, acute (aGvHD) and chronic (cGvHD) Graft-versus-Host Disease were studied taking the competing events of relapse and non-relapse mortality into account.

For illustration, we will concentrate on overall survival and 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 $(1-0.68)-(1-0.66)=-0.02$ 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, due 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 the previous section.

Table 1 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 1 also shows the results from fitting a logistic propensity score model for the probability of receiving an MA conditioning.

Overall survival

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 1 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. Second, we computed pseudo-values at 3 years for each patient based on the Kaplan-Meier estimator. These pseudo-values 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 2.

From either Q-model, the average causal effect on the three-year risk of death was estimated using the G-formula (2) and confidence limits were obtained using the percentile method based on 2000 bootstrap replications. From Table 2 it is seen that the two estimates are quite similar and not very different from the unadjusted estimate based on the Kaplan-Meier estimators. The similarity with the unadjusted value is owing to the low degree of confounding also suggested
Table 1: Pre-treatment characteristics (%) in NM and MA conditioning groups and estimated coefficients in a logistic regression propensity score model.

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>NM (n = 56)</th>
<th>MA (n = 60)</th>
<th>$\hat{\beta}$</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>High cytogeneic risk</td>
<td>25</td>
<td>33</td>
<td>0.404</td>
<td>0.421</td>
</tr>
<tr>
<td>Male donor</td>
<td>54</td>
<td>63</td>
<td>0.465</td>
<td>0.395</td>
</tr>
<tr>
<td>Karnofsky score &lt; 90</td>
<td>14</td>
<td>10</td>
<td>-0.383</td>
<td>0.608</td>
</tr>
<tr>
<td>No prior myelodysplasia</td>
<td>77</td>
<td>82</td>
<td>0.294</td>
<td>0.500</td>
</tr>
<tr>
<td>Sibling donor</td>
<td>68</td>
<td>67</td>
<td>-0.083</td>
<td>0.439</td>
</tr>
<tr>
<td>Same sex donor</td>
<td>57</td>
<td>55</td>
<td>-0.163</td>
<td>0.394</td>
</tr>
</tbody>
</table>

Table 2: Association between pre-treatment characteristics plus MA vs. NM conditioning and overall survival. Left: hazard ratios using a Cox model, right: risk differences using a linear model with pseudo-values as responses. The last line shows the estimated average causal effect (MA vs. NM).

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Cox model</th>
<th>Model for pseudo-values</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>exp($\hat{\beta}$)</td>
<td>$SD(\hat{\beta})$</td>
</tr>
<tr>
<td>MA vs. NM</td>
<td>0.951</td>
<td>0.318</td>
</tr>
<tr>
<td>High cytogeneic risk</td>
<td>1.293</td>
<td>0.339</td>
</tr>
<tr>
<td>Male donor</td>
<td>1.244</td>
<td>0.334</td>
</tr>
<tr>
<td>Karnofsky score &lt; 90</td>
<td>2.192</td>
<td>0.421</td>
</tr>
<tr>
<td>No prior myelodysplasia</td>
<td>1.319</td>
<td>0.419</td>
</tr>
<tr>
<td>Sibling donor</td>
<td>0.618</td>
<td>0.347</td>
</tr>
<tr>
<td>Same sex donor</td>
<td>1.089</td>
<td>0.336</td>
</tr>
<tr>
<td>$ACE$ (MA vs. NM)</td>
<td>-0.013 (-0.181,0.158)</td>
<td>-0.025 (-0.205,0.150)</td>
</tr>
</tbody>
</table>
Table 3: Association between pre-treatment characteristics plus MA vs. NM conditioning and risk of cGvHD. Left: sub-distribution hazard ratios using a Fine-Gray model, right: risk differences using a linear model with pseudo-values as responses. The last line shows the estimated average causal effect (MA vs. NM).

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

by the results in Table 1 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-values were weighted by the inverse probabilities of conditioning treatment allocation with probabilities estimated from the propensity score model in Table 1. 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 weighted data set was -0.018 (-0.166, 0.182). The estimates are close to those obtained using G-estimation.

**Chronic Graft-versus-Host Disease**

A similar set of analyses was carried out in order to estimate the average causal affect of preconditioning treatment on the three-year risk of chronic graft-versus-host disease. Thus, pseudo-values were computed based on the Aalen-Johansen estimator of the cumulative incidence of cGvHD at 3 years taking the competing risks of relapse and non-relapse mortality into account. The $ACE$ was then estimated using, on the one hand, G-estimation and a marginal structural model fitted to an IPT weighted data set both using pseudo-observations and, on the other hand, G-estimation where the Q-model is a Fine-Gray regression model for the cumulative incidence of cGvHD.

Table 3 shows the estimated coefficients from the two Q-models used, i.e., exp($\hat{\beta}$) for the Fine-Gray model and risk differences ($\hat{\beta}$) for the linear model for the pseudo-values 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-values is 0.157 with a 95% c.i. from -0.074 to 0.307.

5 Discussion

We have proposed to use pseudo-values for the purpose of doing causal inference in survival analysis. Since pseudo-values provide a general tool for analyzing mean value parameters in the presence of right-censoring, the method lends itself for 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 (Andersen et al., 2004) or the number of years spent in a given state in a multi-state model (Grand and Putter, 2016; Andersen, 2013). 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. (2014) showed how alternative IPCW estimators may form the basis for computing the pseudo-values 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 $I(Y_i \leq t, D_i = 1)$.

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

In our example the impact of using pseudo-values 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 1). 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 above, albeit with considerably 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 to 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-value approach avoids such assumptions.

Software for calculating pseudo-values are available in both SAS and R (Klein et al., 2008) and in Stata (Parner and Andersen, 2010; Overgaard et al., 2015).

References


Austin, P.C. (2010). Absolute risk reductions and numbers needed to treat can be obtained from adjusted survival models for time-to-event outcomes. Journal of Clinical Epidemiology 63, 46-55.


Austin, P.C. (2014). 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 33, 1242-1258.


Research Reports available from Department of Biostatistics

http://www.pubhealth.ku.dk/bs/publikationer

Department of Biostatistics
University of Copenhagen
Øster Farimagsgade 5
P.O. Box 2099
1014 Copenhagen K
Denmark

13/01 Forman, J.L. & Sørensen M. A new approach to multi-modal diffusions with applications to protein folding.


13/03 Wolbers, M., Koller, M.T., Witteman, J.C.M. & Gerds, T.A. Concordance for prognostic models with competing risks.

13/04 Andersen, P.K., Canudas-Romo, V. & Keiding, N. Cause-specific measures of life lost.

13/05 Keiding, N. Event history analysis.


13/07 Hansen, S., Andersen, P.K. & Parner, E. Events per variable for risk differences and relative risks using pseudo-observations.


13/11 Olsbjerg, M. & Christensen K.B. SAS macro for marginal maximum likelihood estimation in longitudinal polytomous Rasch models.

14/01 Ambrogi, F. & Andersen, P.K. Predicting Smooth Survival Curves through Pseudo-Values.

14/02 Olsbjerg, M. & Christensen, K.B. Modeling local dependence in longitudinal IRT models.

14/03 Olsbjerg, M. & Christensen, K.B. LIRT: SAS macros for longitudinal IRT models.

14/04 Jacobsen, M. & Martinussen, T. A note on the large sample properties of estimators based on generalized linear models for correlated pseudo-observations.

14/05 De Neve, J. & Gerds, T. A note on the interpretation of the Cox regression model.

14/06 Gerds, TA. The Kaplan-Meier theater.

14/07 Martinussen, T., Holst, K.K. & Scheike, T. Cox regression with missing covariate data using a modified partial likelihood method.

15/01 Mansourvar, Z., Martinussen, T. & Scheike, T. Semiparametric regression for restricted mean residual life under right censoring.

15/02 Mansourvar, Z., Martinussen, T. & Scheike, T. An additive-multiplicative restricted mean residual life model.

15/03 Mansourvar, Z. & Martinussen, T. Estimation of average causal effect using the restricted mean residual lifetime as effect measure.


15/06 Müller, M. & Kreiner, S. Item Fit Statistics in Common Software for Rasch Analysis.

16/01 Andersen, P.K. Life years lost among patients with a given disease.

16/02 Strohmaier, S., Haase, N., Wetterslev, J. & Lange, T. A simple to implement algorithm for natural direct and indirect effects in survival studies with a repeatedly measured mediator.

16/03 Andersen, P.K., Syriopoulou, E. & Parner, E. Causal inference in survival analysis using pseudo-observations.