Biostatistics (2019), **0**, 0, *pp.* 1–18
doi:10.1093/biostatistics/output

Causal inference in continuous time: An example on prostate cancer therapy

PÅL CHRISTIE RYALEN*

Department of Biostatistics, University of Oslo, Domus Medica Gaustad, Sognsvannsveien 9, 0372 Oslo, Norway

MATS JULIUS STENSRUD

Department of Biostatistics, University of Oslo, Domus Medica Gaustad, Sognsvannsveien 9, 0372 Oslo, Norway
Diakonhjemmet hospital, Department of Medicine, Diakonveien 12, 0370, Oslo, Norway

SOPHIE FOSSÅ

National Advisory Unit on Late Effects after Cancer Treatment, Oslo University Hospital, Radiumhospitalet, 0424 Oslo

Institute of Clinical Medicine, Oslo University Hospital, Søsterhjemmet, Kirkeveien 166, 0450 Oslo, Norway

Cancer Registry of Norway, Ullernchausseen 64, 0379 Oslo, Norway

KJETIL RØYSLAND

Department of Biostatistics, University of Oslo, Domus Medica Gaustad, Sognsvannsveien 9, 0372 Oslo, Norway

p.c.ryalen@medisin.uio.no

Summary

In Marginal Structural Models (MSMs), time is traditionally treated as a discrete parameter. In survival analysis on the other hand, we study processes that develop in continuous time. Therefore, Røysland (2011) developed the continuous-time MSMs, along with continuous-time weights. The continuous-time weights are conceptually similar to the inverse probability weights that are used in discrete time MSMs. Here we demonstrate that continuous-time MSMs may be used in practice. First, we briefly describe the causal model assumptions using counting process notation, and we suggest how causal effect estimates can be derived by calculating continuous-time weights. Then, we describe how additive hazard models can be used to find such effect estimates. Finally, we apply this strategy to compare medium to long term differences between the two prostate cancer treatments Radical prostatectomy (RP) and Radiation therapy (Rad), using data from the Norwegian Cancer Registry. In contrast to the results of a naive analysis, we find that the marginal cumulative incidence of treatment failure is similar between the strategies, accounting

*To whom correspondence should be addressed.

for the competing risk of other death.

Key words: Causal inference in survival analysis;Continuous-time Marginal Structural Models;Continuous-time weights;Prostate cancer therapy.

1. Introduction

To evaluate the effect of medical interventions, randomized controlled trials (RCTs) are desirable. Such trials, however, are often lacking. RCTs are hard to conduct in many real-life settings, because they tend to be expensive, time-consuming and sometimes unethical.

In contrast, relevant observational data are often available. Nowadays individual patient records are frequently collected in clinical registries. These registries may be essential to decision makers who can benefit from data that are cheap to analyze. On the other hand, analyses of observational data are fundamentally prone to confounding and selection bias. To seriously address these concerns, there have been major methodological developments in the last decades. Strategies for causal inference in observational studies have been improved, and several methods are now accessible for a broader range of researchers. By explicitly stating the causal assumptions, such methods allow us to interpret the findings from observational studies causally. In this framework, observational studies are commonly designed to mimic randomized controlled trials.

In many medical scenarios, the outcome of interest is the time to an event. In particular, we may be interested in how cancer treatments, e.g. a surgical intervention and radiation, affect survival times. To evaluate time to event outcomes, we need to observe subjects over time. The longitudinal structure may complicate the analysis, because the effects may e.g. be time-dependent and subjects may drop out from the study.

To account for time-varying confounding, Robins (1997) developed the marginal structural models. The strategy requires the assumption of no unmeasured confounding, and, heuristically, causal estimates are derived by weighting the subjects to create a pseudopopulation unaffected by confounding and selection bias. By using weighting to control for confounding, MSMs allow for flexible definitions of the structural model, which is not only desirable when time varying confounding is present, but also in other scenarios (Joffe *and others*, 2004). For time to event exposures, however, this strategy has considered time to be a discrete unit (Hernán, Brumback and Robins, 2002). In this approach, IPT weights are typically estimated using pooled logistic regression where the event times are binned into discrete intervals. This binning of event times is known to induce either bias or high variance (Ryalen, Stensrud and Røysland, 2018). Since time to event outcomes usually occur in continuous time, we aim to develop method that consider time to be a continuous process.

In the probability literature, Røysland (2011) described MSMs in continuous time using martingale theory. The idea is that desirable features such as randomization can be framed as a change of probability measure, from an observational measure to a hypothetical measure. The two measures are related through a Radon-Nikodym derivative that serves as a weight process. By weighting the observed data, we can mimic observations that are derived from the hypothetical measure. These methods are theoretically sound and applicable to various observational survival data, but they have not been applied in the literature.

In this article, we apply continuous-time weights to compare two prostate cancer treatment regimens, radiation (Rad) and radical prostatectomy (RP), using data from the Norwegian cancer registry. Our outcome of interest is time to failure of treatment, which is a surrogate for death due to prostate cancer. There may be confounding between the outcome and treatment assignment,

and there may be confounding between the outcome and the time to treatment initiation. To account for the confounding, we attempt to study a scenario where both the treatment mode (Rad or RP) and the time to treatment initiation are randomized. To construct this scenario, we rely on propensity weights for the treatment mode, and continuous-time weights for time to treatment initiation. To assess the outcome, we describe a marginal structural cumulative incidence model as a function of the two treatment modalities. The model parameters are the cumulative incidences of death due to prostate cancer if, contrary to fact, we had imposed Rad treatment on the entire population versus if we had imposed RP treatment, and under either regime, had we ensured that the treatment initiation rates were randomized. We obtain weighted nonparametric plug-in estimators that are known to be consistent, assuming no unmeasured confounders, exchangeability and that the model is correctly specified. Indeed, this strategy is a general three-step approach to causal survival analysis: 1) Estimate continuous-time weights using additive hazard models, where covariate selection is informed by local independence graphs. 2) Fit weighted additive hazard regressions for the time to event outcome. 3) Transform the weighted hazard estimates to obtain parameters that allow for a causal interpretation (Ryalen, Stensrud and Røysland, 2017, 2018).

The remainder of this paper is organized as follows. In Section 2 we give a brief account of the theory that is necessary for our causal analysis. Then we describe sources of bias in the registry data, and we suggest how to adjust for bias by weighting. Next we describe an estimation procedure for the continuous-time weights, propose a MSM, before explaining a method for testing equality of hypothetical cumulative incidences. In Section 3 we give details regarding model fitting and the weighted analysis, and give outcomes of the performed tests. A discussion is provided in Section 4.

2. Deriving causal estimates

Our definition of a causal counting process model is derived from Røysland (2011), but we include an explicit formulation to clarify the concept. Throughout the manuscript we will use subscript notation to indicate the value of a process at time t.

2.1 *Representation of patient information*

We consider n i.i.d. individuals over a study period $[0, \mathcal{T}]$. Each individual i is represented by a set of d random variables $X^{i,1}, \dots, X^{i,d}$ at baseline, and k counting processes $N^{i,1}, \dots, N^{i,k}$. In our example, $N_t^{i,j}$ for a particular j will e.g. denote whether or not a particular prostate cancer treatment is initiated by time t for patient i. In other examples, $N_t^{i,j}$ could also be a time-dependent confounder. The filtration $\mathcal{F}_t$ is generated by all measurable events for every individual that can occur before time t. In our example, $\mathcal{F}_t$ will e.g. include information on all the baseline variables as well as the treatment initiation times and the treatment failure times until t. Let P denote the joint density that governs the frequencies of all the events in $\mathcal{F}_\mathcal{T}$. Moreover, let $\lambda^{i,j}$ denote the intensity for the counting process $N^{i,j}$ with respect to P and $\mathcal{F}_\mathcal{T}$. Heuristically, this means that

$$\lambda_t^{i,j} = \lim_{\Delta \to 0} \frac{1}{\Delta} P\big(N_t^{i,j} - N_{t-\Delta}^{i,j} > 0 | \mathcal{F}_{t-\Delta}\big), \tag{2.1}$$

so the intensity can be interpreted as a conditional density in a short time interval. From (Jacod, 1975) we know that the density of the counting processes conditional on the baseline variables are uniquely characterized by the intensities. We also know that P at baseline is given by the

conditional densities

$$dP\big(X^{i,m}|X^{i,m-1},\ldots,X^{i,1}\big), \tag{2.2}$$

for $m = 1,\cdots,d$, as the joint density at baseline factorizes as a product of conditional densities. We can thus determine P uniquely knowing the factors (2.1) and (2.2).

2.2 *Causal validity*

Suppose that each $N^{i,j}$ represents an exposure for individual i that in principle would allow a hypothetical, but meaningful, intervention γ that would change the joint measure for all the events to a hypothetical measure $\tilde{P}$: The intervention would impose a $\tilde{P}$ intensity $\tilde{\lambda}^{i,j}$ for $N^{i,j}$. If the intervention is targeted at the j'th process $N^{i,j}$, that e.g. represents the treatment trajectory, then this intervention would simply correspond to changing treatment regime $\lambda^{i,j}$ into $\tilde{\lambda}^{i,j}$. In our example, we aim to define a hypothetical treatment strategy in which (i) RP and Rad are initiated independent of baseline covariates, and (ii) the time-to-treatment is equally distributed for both treatment regimens. To do this, we will consider a scenario in which the time to treatment initiation is derived from a marginal hazard model for the cohort as a whole, regardless of treatment regimen.

Consider the intervention γ that imposes the $\tilde{P}$ intensity $\tilde{\lambda}^{i,j}$ on $N^{i,j}$, the j'th counting process of individual i. Our model is said to be **causal** if applying γ would not have changed the 'local characteristics' for the other components . More precisely this means that

- The functional $\lambda^{i,m}$ would also define the intensity for $N^{i,m}$ with respect to $\tilde{P}$ when $m \neq j$, and

- The conditional densities in (2.2) would be the same with respect to both P and $\tilde{P}$, i.e.

  $$dP\big(X^{i,m}|X^{i,m-1},\ldots,X^{i,1}\big) = d\tilde{P}\big(X^{i,m}|X^{i,m-1},\ldots,X^{i,1}\big),$$

  for $m = 1,\cdots,d$

If the intervention instead was targeted at a baseline-variable, say $X^{i,j}$, and this intervention would replace the conditional densities in (2.2) by densities on the form $d\tilde{P}\big(X^{i,m}|X^{i,m-1},\ldots,X^{i,1}\big)$, for $m = 1,\cdots,d$, then the model is said to be causal if

- The functional form of the intensity process of $N^{i,m}$ with respect to P and $\tilde{P}$ coincide for $m = 1,\cdots,k$, and

- The remaining conditional densities at baseline coincide, i.e.

  $$dP\big(X^{i,m}|X^{i,m-1},\ldots,X^{i,1}\big) = d\tilde{P}\big(X^{i,m}|X^{i,m-1},\ldots,X^{i,1}\big),$$

  for $m \neq j$.

Note that the formulation of a baseline intervention agrees with Pearl's definition of a causal model (Pearl, 2000). Furthermore, conditional independence in a Bayesian network corresponds to local independence in our time-continuous setting.

2.3 *The weights*

For the baseline intervention, a basic measure theoretic argument shows that if $\tilde{P} \ll P$, then the likelihood ratio from P to $\tilde{P}$ for individual i is given by:

$$W_0^i := \frac{d\tilde{P}\big(X^{i,j}|X^{i,j-1},\ldots,X^{i,1}\big)}{dP\big(X^{i,j}|X^{i,j-1},\ldots,X^{i,1}\big)}. \tag{2.3}$$

This is simply the ordinary propensity-score weight, see Pearl (2000).

For an intervention aimed at one of the counting-processes, the likelihood ratio for individual i at time t is given by the so called Jacod-formula (Jacod, 1975, (14)), that reads

$$W_t^i = \prod_{s \leqslant t} (\theta_s^{i,j})^{\Delta N_s^{i,j}} e^{\int_0^t \lambda_u^{i,j} - \tilde{\lambda}_u^{i,j} du}, \tag{2.4}$$

where $\theta^{i,j} = \tilde{\lambda}^{i,j}/\lambda^{i,j}$. If the intervention is targeted at a process $N_t^{i,j}$ that counts treatment initiation, individual i's weight W_t^i process will change in a continuous manner whenever he is at risk of being treated. If he receives treatment at time τ, then $\Delta N_\tau^{i,j} = 1$ and his weight process will jump by a factor $\theta_\tau^{i,j}$. Whenever he is not at risk of receiving treatment, the weight process will be constant.

If we knew W_t^i we could re-weight the observed data to estimate parameters that would be obtained if we had carried out our hypothetical intervention γ. However, these formulas only represent the theoretical weights that are likely to be unknown in most statistical situations: In real life, we have to find good estimates of these weights, usually using the data at hand.

We will use the fact that (2.4) solves the Doleans-Dade equation. This allows us to write it as a simple integral equation with separate parts driven by the counting process $N^{i,j}$ and the cumulative intensity terms $\Lambda_t^{i,j} = \int_0^t \lambda_s^{i,j} ds$ and $\tilde{\Lambda}_t^{i,j} = \int_0^t \tilde{\lambda}_s^{i,j} ds$:

$$W_t^i = 1 + \int_0^t W_{s-}^i (\theta_s^{i,j} - 1) dN_s^{i,j} + \int_0^t W_s^i d\Lambda_s^{i,j} - \int_0^t W_s^i d\tilde{\Lambda}_s^{i,j}. \tag{2.5}$$

2.4 *Calculating continuous-time weights*

We consider longitudinal data in which individual i's time from, say, diagnosis to treatment is governed by the intensity $\lambda^{i,j}$. We are interested in making inference in a world where the treatment regime is governed by an intensity $\tilde{\lambda}^{i,j}$. In many situations we may be interested in outcomes of (hypothetical) randomized trials, and $\tilde{\lambda}^{i,j}$ will be a marginal intensity, in our scenario this marginal intensity is derived from the pooled treatment initiation rate in the population. Suppose that we can find reasonable hazard models for the treatment initiation times in the sample (in our example, e.g. an additive model describing time to treatment initiation). Hazard models typically give cumulative hazards estimates; that is, estimates of the integrated hazard as a function of time, $A_t = \int_0^t \alpha_s ds$, where α is a hazard rate. Using the fitted hazard models we can extract cumulative hazard estimates $\hat{A}^{i,j}$ and thus obtain cumulative intensity estimates by using the multiplicative intensity model $\hat{\Lambda}_t^{i,j} = \int_0^t Y_s^{i,j} d\hat{A}_s^{i,j}$, where $Y^{i,j}$ is the at-risk-for-treatment indicator. From the estimated cumulative intensities, as well as some estimate $\hat{\theta}^{i,j}$ of $\theta^{i,j}$, we can estimate W^i by simply plugging $\hat{\Lambda}^{i,j}, \hat{\tilde{\Lambda}}^{i,j}$ and $\hat{\theta}^{i,j}$ into (2.5), obtaining

$$\hat{W}_t^i = 1 + \int_0^t \hat{W}_{s-}^i (\hat{\theta}_{s-}^{i,j} - 1) dN_s^{i,j} + \int_0^t \hat{W}_{s-}^i d\hat{\Lambda}_s^{i,j} - \int_0^t \hat{W}_{s-}^i d\hat{\tilde{\Lambda}}_s^{i,j}. \tag{2.6}$$

The counting process term $\int_0^t \hat{W}^i_{s-}(\hat{\theta}^{i,j}_{s-} - 1)dN^{i,j}_s$ will only make a contribution when individual i has an event, e.g. receiving treatment. Our candidate for estimating the intensity ratio is

$$\hat{\theta}^{i,j}_t = \frac{\Delta_b \hat{\tilde{\Lambda}}^{i,j}_t}{\Delta_b \hat{\Lambda}^{i,j}_t}, \tag{2.7}$$

where Δ_b is a difference operator defined by $\Delta_b D_t = D_t - D_{t-b}$, for a suitable smoothing parameter b. In practice, $\tilde{\Lambda}^{i,j}$, and hence the numerator of (2.7), is typically estimated from a reference population, even though this process could have been entirely hypothetical.

The estimator (2.7) can give extreme values when the denominator is close to zero, resulting in large or even negative estimates. The denominator being close to zero is a violation of the positivity condition, as the estimated weights then fail to approximate the true likelihood ratio. Such weights will play a huge role when performing weighted analyses, and thus the weight estimation must be treated with care. In our experience extreme weights has occurred as a result of a misspecified model, or for individuals that are fundamentally different from the rest of the cohort, i.e. for outlying individuals. This is similar to violation of positivity for e.g. propensity models, but we stress that the choice of model parametrization is likely to be very important for the continuous-time weights.

We have implemented weight estimators based on the additive hazard model (Aalen, 1989) in the `R` package `ahw`. Equation (2.6) is easy to solve on a computer in this case, since the integrals on the right hand side then can be reduced to counting process integrals. The estimator (2.6) then reduces to a piecewise constant, recursive equation, and can easily be solved on a computer using e.g. a `for` loop. We have used the `ahw` package to perform the weighted analysis in this article.

2.5 *The clinical question and sources of bias*

We follow patients from time of diagnosis, and we register when (i) treatment is received, and (ii) death or end of follow-up occurs. We aim to compare the effectiveness of the treatment regimens RP and Rad. Due to the lack of randomization, we must consider two major sources of bias. First, the assigned treatment regimen varies among subgroups of the population. In our scenario, patients receiving RP treatment are e.g. almost four years younger than patients in the Rad group on average, and this is also consistent with previous trials (Krupski *and others*, 2005). Socioeconomic factors such as income may also influence the choice of treatment (Krupski *and others*, 2005; Woods, Rachet and Coleman, 2005). Second, the rate at which individuals start treatment depends on the treatment regimen, as well as individual characteristics such as education level, Gleason score, and Prostate specific antigen (PSA) levels. In an ideal randomized trial we would not have such dependencies, and the treatment groups would be comparable at the time of treatment.

2.6 *Mimicking randomization by weighting*

To adjust for these sources of bias, we will rely on a weighted analysis. We assume that the important causal relationships are incorporated in the local independence graph displayed in Figure 1, which describe conditional dependencies (Readers unfamiliar with local independence graphs may find Didelez (2008) useful).

We assume that age, education and diagnostic measures such as PSA and Gleason score may influence both the treatment regimen, the time at which treatment is initiated and the outcome,

e.g. death from prostate cancer. In a desired RCT, the treatment regimen is selected randomly, and the rate of treatment initiation is the same for all subjects. Removing the arrow from the diagnostic factors to Treat mode, and the arrow from Treat mode to Treat start in Figure 1 gives us a local independence diagram that would describe the dynamics in such a trial. If we assume 'no unmeasured confounding', i.e. that this model is causal, we get that any association between Treat mode and Failure would represent a causal effect.

Our hypothetical randomization strategy involves two interventions - one on the Treat mode at baseline, and one on the treatment initiation counting process. From Section 2.3, under causal validity, we know that the frequency of all the events in the hypothetical scenario would be governed by the actually observed frequencies, re-weighted by baseline weights (2.3) and continuous-time weights (2.5). Thus the weighting can be partitioned into two steps. First, we calculate inverse probabilities of the choice of treatment at baseline by fitting logistic regression models. Second, to mimic randomization of the time from diagnosis to treatment, we will employ continuous-time treatment weights. As described in Section 2.3, we find these weights by creating a treatment intensity $\tilde{\lambda}$ that corresponds to the desired randomized experiment. This intensity will be marginal, i.e. the intensity is the same for all individuals regardless of variables like treatment group, PSA levels or age. We let $\tilde{\lambda}$ be the marginal treatment initiation intensity for the pooled observed data. In this way, each individual receives a propensity weight at baseline, and a weight process from the time of diagnosis until end of follow-up. Re-weighting the data according the product of the propensity weights and the continuous weight-process provides the observational frequencies we would see if we randomized to correct for (i) confounding of the choice of treatment, and (ii) the time to treatment initiation.

2.7 *Competing risks*

The treated individuals are in a competing risk situation between treatment failure and death by other causes. We therefore introduce two terminating endpoints; Failure and Other death (see Supplementary Fig. 2 for a multi-state representation). We will evaluate cumulative incidences, i.e. comparing the cumulative risk of Failure as a function of time, allowing for the competing event of Other death to occur. The interesting question is whether the cumulative incidences of Failure differ between the RP group and the Rad group, and whether such a difference is due to the choice of treatment rather than due to covariate imbalances between the treatment groups. Thus, our estimands of interest are the cumulative incidences of Failure that we would observe if, contrary to fact, each treatment had been imposed on all subjects in the population, and under each treatment, the time to treatment initiation was a random draw from the marginal distribution of the treatment initiation times in the sample. These estimands would be easy to calculate using standard survival analysis techniques if the data came from the desired randomized trial. If T_f is the time from diagnosis to treatment failure, T_{od} denotes the time to other death, and the observational scenario P actually were the randomized trial, the estimands would simply be estimated by the cumulative incidence of Failure within each group; $P(t \geqslant T_f, T_f < T_{od}|l)$ for $l \in \{\text{Rad}, \text{RP}\}$.

2.8 *A marginal structural model*

We aim to use our observational data to mimic a randomized trial. Let g_l denote an intervention where treatment mode is set to l while the treatment is initiated according to the marginal initiation rate in the registry data. The distribution of events under this intervention is denoted

by $\tilde{P}^{g_l}$. A saturated non-parametric marginal structural model for the cumulative incidence under treatment regime l is then

$$\begin{aligned} G_t^l &= \tilde{P}^{g_{\text{Rad}}}(t \geqslant T_f, T_f < T_{od}) I(l = \text{Rad}) + \tilde{P}^{g_{\text{RP}}}(t \geqslant T_f, T_f < T_{od}) I(l = \text{RP}) \\ &= \tilde{C}_t^{\text{Rad}} I(l = \text{Rad}) + \tilde{C}_t^{\text{RP}} I(l = \text{RP}) \end{aligned} \tag{2.8}$$

Assuming that the treatment groups are exchangeable, we have that $\tilde{P}^{g_l}(t \geqslant T_f, T_f < T_{od}) = \tilde{P}^{g_l}(t \geqslant T_f, T_f < T_{od} | l)$ for each l. By assuming exchangeability, positivity and no unmeasured confounding, we can identify the coefficients $\tilde{C}^{\text{Rad}}$ and $\tilde{C}^{\text{RP}}$ in (2.8) by performing a weighted analysis within each of the treatment groups. We have now obtained a non-parametric structural model that is easy to estimate: After weighting the original data, we may simply study cumulative incidence curves, e.g. derived by Nelson-Aalen estimators, for each of the treatment groups. We only rely on parametric assumptions for the weight estimation.

2.9 *Estimating, and testing equality of hypothetical cumulative incidences*

We estimate the hypothetical cumulative incidences $\tilde{C}^l$, the cumulative incidences we would have observed under our hypothetical randomization, by recognizing that they solve ordinary differential equation systems on the form

$$\begin{pmatrix} \tilde{C}_t^l \\ \tilde{S}_t^l \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \end{pmatrix} + \int_0^t \begin{pmatrix} \tilde{S}_s^l & 0 \\ -\tilde{S}_s^l & -\tilde{S}_s^l \end{pmatrix} d \begin{pmatrix} \tilde{A}_s^{l,\text{Failure}} \\ \tilde{A}_s^{l,\text{Other death}} \end{pmatrix},$$

for group $l \in \{\text{Rad}, \text{RP}\}$. Here, $\tilde{A}^{l,\text{Failure}}$ and $\tilde{A}^{l,\text{Other death}}$ are the hypothetical cumulative hazards (i.e. cumulative hazards under P^{g_l}) for the transition from state l to the Failure and Other death endpoints, respectively, while $\tilde{S}^l$ is the hypothetical survival function (i.e. the survival function under P^{g_l}). Our estimator is obtained by simply "plugging in" estimates of the integrators, which will be the weighted cause-specific Nelson-Aalen estimators $\hat{A}^{l,\text{Failure}}, \hat{A}^{l,\text{Other death}}$. The "weighting" is simply performed by multiplying the i'th at risk indicator at time t with $\hat{W}_{t-}^i$ for each $i = 1, \cdots, n$ in the standard Nelson-Aalen estimator. From Section 2.6 we recall that $\hat{W}^i$ in our case is the product of the estimated propensity weight and continuous-time weight for individual i. We end up with a naturally associated system that reads

$$\begin{pmatrix} \hat{\tilde{C}}_t^l \\ \hat{\tilde{S}}_t^l \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \end{pmatrix} + \int_0^t \begin{pmatrix} \hat{\tilde{S}}_{s-}^l & 0 \\ -\hat{\tilde{S}}_{s-}^l & -\hat{\tilde{S}}_{s-}^l \end{pmatrix} d \begin{pmatrix} \hat{\tilde{A}}_s^{l,\text{Failure}} \\ \hat{\tilde{A}}_s^{l,\text{Other death}} \end{pmatrix}. \tag{2.9}$$

This is a recursive, piecewise constant equation system that is similar to (2.6).

We can test for equality of the cumulative incidences using the approach suggested by Gray (1988), i.e. a log-rank test based on comparing subdistribution hazards. We are, however, interested in (hypothetical) randomized trial outcomes, and will consequently compare weighted subdistribution hazards. By plugging in the weighted quantities the test statistic reads

$$\hat{\tilde{Z}}_{\mathcal{T}} = \int_0^{\mathcal{T}} K_t \cdot \Big(\frac{d\hat{\tilde{C}}_t^{\text{RAD}}}{1 - \hat{\tilde{C}}_{t-}^{\text{RAD}}} - \frac{d\hat{\tilde{C}}_t^{\text{RP}}}{1 - \hat{\tilde{C}}_{t-}^{\text{RP}}} \Big), \tag{2.10}$$

for a suitable function K.

The estimation procedure that gave rise to (2.9) applies to several other survival analysis quantities, e.g. relative survival and restricted mean survival, which may be useful for causal

survival analysis (Ryalen, Stensrud and Røysland, 2017, 2018). Assuming that the cumulative hazards follow an additive model (Aalen, 1989), and using the weight estimator proposed in Section 2.4, such plug-in estimators are consistent.

Thus we can state that the quantities (2.9) and (2.10) are consistent. In fact, the latter is asymptotically normally distributed with mean zero under the hypothesis that the hypothetical subdistribution hazards $\tilde{\zeta}^l_t dt = \frac{d\tilde{C}^l_t}{1-\tilde{C}^l_t}$ are equal on the study period.

3. Model fitting and analysis

We considered Norwegian males from the cohort diagnosed in 2004-2005, registered by the Norwegian cancer registry until 2015. We restricted our analysis to nonmetastatic subjects who were assigned to a treatment group; Rad or RP. Subjects who did not receive treatment, e.g. due to serious health problems, were not included in the analysis. Thereby we selected subjects who got treatment, which in principle may lead to selection bias. However, the fraction lost from treatment assignment to treatment initiation is likely to be minor, due to the short time period, and the fact that only those who are likely to receive treatment are assigned to treatment. Also, we left out individuals in the highest risk group, because these individuals will in practice always receive radiation therapy. Hence, we restricted our analysis to subject who, according to current clinical practice, would be eligible to both treatment strategies.

A summary of the baseline characteristics is found in Table 1, and more information about the data is found in the supplementary material.

As explained in Section 2.6, we fit logistic regression models for the propensity weights; a marginal model for the weight numerator, and a covariate dependent model for the denominator. In the denominator we modeled the probability that individual i receive RP treatment by

$$\begin{aligned}\text{logit}(p_i) = p_0 &+ p_{\text{PSA}_{(8,15]}} I(\text{PSA}_i \in (8,15]) + p_{\text{PSA}_{(15,22]}} I(\text{PSA}_i \in (15,22]) + p_{\text{age}>65} I(\text{age}_i > 65) \\ &+ p_{\text{CAD}} I(\text{CAD}_i = 1) + p_{\text{HYP}} I(\text{HYP}_i = 1) + p_{\text{earlier cancer}} I(\text{earlier cancer}_i = 1) \\ &+ p_{\text{gleason}>6} I(\text{gleason}_i > 6) + p_{\text{T cat}} I(\text{T cat}_i = 1) + p_{< \text{ high school}} I(\text{edu}_i < \text{ high school}) \\ &+ p_{>4 \text{ years college}} I(\text{edu}_i > 4 \text{ years college}) + p_{\text{risk group}} I(\text{risk group}_i > 1),\end{aligned}$$

where CAD is Coronary artery disease and HYP is Hypertension, risk group is a three-valued risk variable that is a combination of PSA, Gleason and T cat. It takes the values 1(low risk), 2(intermediate risk) and 3(high risk). For the continuous-time weights we fit additive hazard models for time to treatment. The patients are split into the two treatment groups, and we fit covariate dependent models for each of the groups on the form:

$$\begin{aligned}\alpha^i_t = \alpha^0_t &+ \alpha^{\text{age}>65}_t I(\text{age}_i > 65) + \alpha^{\text{PSA}>5}_t I(\text{PSA}_i > 5) + \alpha^{\text{gleason}>6}_t I(\text{gleason}_i > 6) \\ &+ \alpha^{< \text{ high school}}_t I(\text{edu}_i < \text{ high school}) + \alpha^{>4 \text{ years college}}_t I(\text{edu}_i > 4 \text{ years college}) \\ &+ \alpha^{\text{risk group}}_t I(\text{risk group}_i > 1).\end{aligned}$$

We repeat that the model is fitted for the Rad group and RP group separately. We also fit a marginal additive hazard model $\tilde{\alpha}^i_t$ for time to treatment initiation in the pooled sample. Following Section 2.6 it is the same for all individuals regardless of treatment group, thus $\tilde{\alpha}^i = \tilde{\alpha}$. Since it is marginal, we can estimate the cumulative hazard using a Nelson-Aalen estimator. From the additive hazard regression we obtain cumulative hazard estimates, allowing us to calculate the continuous time weights in the manner described in Section 2.4. Further details on the model fitting, including numerical estimates on the coefficients and diagnostic plots, are found in the Supplementary material.

3.1 *Defining the failure endpoint*

Ideally, our primary outcome would have been death by prostate cancer. However, obtaining satisfactory power with this outcome would require a longer follow-up time and a larger cohort. To increase the power, we therefore consider the composite endpoint 'Failure of curative treatment' or 'Failure', i.e. the first registered indication that the curative treatment was unsuccessful. We therefore defined Failure to be the first occurrence of either

- For RP patients: Radiation therapy later than 6 months after initial treatment.
- For Rad patients: New radiation treatment applied later than 8 weeks after initial treatment.
- End of a 6 months gap or more of hormone treatment.
- Continued hormone treatment later than 3 years after initial treatment.
- Death by prostate cancer.

Our outcome is a surrogate endpoint, and may be justified as follows: First, RP treatment is intended to be curative, and should not require subsequent radiation therapy. Any radiation treatment given prior to 6 months is viewed as adjuvant. Therefore, radiation therapy after 6 months in RP patients indicates that the initial treatment was unsuccessful. Likewise, the initial radiation therapy treatment failed if a subsequent radiation therapy course was started, usually lasting for eight weeks, the expected time of the curative radiation cycle.

Third, Rad patients receive adjuvant hormone treatment for 3 years. A longer period of hormone therapy indicates that there are further signs of disease. Some patients also stop hormone treatment before three years, possibly because of side effects. If hormone treatment is restarted after some time, we consider the initial curative treatment to be unsuccessful.

3.2 *Weighted analysis*

Cumulative incidences under the two treatment regimens are shown in Figure 2. The difference between the weighted cumulative incidences, i.e. estimates of $G^{\mathrm{Rad}} - G^{\mathrm{RP}}$, along with bootstrapped confidence intervals are shown in Figure 3. The failure rates seem to be similar when mimicking a randomized trial. In the Supplementary material, we have also performed a discrete time MSM analysis using pooled logistic regression, which resulted in similar (but not identical) curves.

More rigorously, we perform the weighted Gray test (2.10). The differences between the curves that occur shortly after diagnosis are likely to be a result of our particular choice of Failure endpoint: We have described particular time intervals, e.g. a 6 months gap on hormone treatment or a application of radiation therapy after 8 weeks, as events. These events may be sensitive to the particular definition of the outcome, e.g. the choice of a gap of 6 months for RP treatment, instead of, say, 4 months. However, our target of this analysis is the longer time effects of treatment. We therefore used the function $K_t = t^{0.3}$ in (2.10), putting less emphasis on differences between the weighted curves for small t. By bootstrapping the variance with a bootstrap sample of 3000 we get a p-value of 0.3310, indicating that the hypothetical cumulative incidences are not very different. A similar test for the unweighted cumulative incidences gave a p-value of 0.0083, and a naive conclusion may have been that Rad treated individuals had a significantly worse failure rate. The functions $K_t = t^{0.2}$ and $K_t = t^{0.5}$ were also used, yielding the same conclusions.

4. Discussion

We have shown how non-randomized clinical data can be used in a causal survival analysis. In contrast to previous causal approaches using MSMs, we have treated time as a continuous variable, not (artificially) discrete. While we have not considered time-dependent confounding in this article, the continuous-time approach also deals with such scenarios (Røysland, 2011). In our example, the time from diagnosis to treatment initiation was relatively short (median 137 days, see the Supplementary materials), and data on time-updated measurements were not available. However, despite the short time from diagnosis to treatment initiation, we cannot be certain that some covariates influencing both the exposure and the outcome are time-varying in this period, potentially leading to time-dependent confounding.

Nevertheless, we believe that MSMs are useful in many survival analysis settings, not only in scenarios with time-dependent confounding: Heuristically, by obtaining balanced treatment groups through continuous-time weighting, we are flexible to define a structural model for the outcome estimand, which in our scenario involved hypothetical cumulative incidences. Furthermore, the weighting approach allows us to study the appropriate time scale of interest: We obtain the treatment effect measured from the time of diagnosis, which we believe is the most relevant to patients and decision makers. Other approaches, e.g. use of propensity weighted regression, may have allowed us to identify the effect measured from the time of treatment initiation, but it would not be trivial to account for the time between diagnosis and treatment initiation. In principle, adjusting for differences in time to treatment may be done in an outcome regression model if there is no time-varying confounding. However, such modelling would require an appropriate parametrization of how time from diagnosis to treatment affects the outcome, which is far from trivial in practice.

A different, but more common alternative to our weighted survival analysis is to assume a proportional hazards model. The hazard scale is convenient and flexible for regression modeling, but estimates on the hazard scale, e.g. hazard ratios, often lack a clear causal interpretation (Hernán, 2010). The problem arise because conditioning on recent survival can open non-causal pathways, and will therefore make causal interpretations less obvious in many situations. Our approach using cumulative incidences, which has also been the target parameter in discrete time MSMs (Moodie, Stephens and Klein, 2014), do not suffer from this bias. Thereby, we do not need to rely on the proportional hazards assumption.

We suggest that causal survival analysis may be performed in three steps: First, assuming a causal intervention can be applied, hazard scale regression may be used to estimate treatment weights. Second, weighted outcome hazard regressions can be fitted to obtain causal cumulative hazard estimates. Third, cumulative hazard estimates can be transformed to estimate other parameters by use of differential equations. Steps one and two are described in (Ryalen, Stensrud and Røysland, 2018), while step three is shown in (Ryalen, Stensrud and Røysland, 2017). This allows us to study causal parameters in survival analysis that are easy to interpret. Currently we rely on bootstrapping for inference, but developing analytic strategies to estimate the uncertainty is an important topic for future research.

We emphasise that the continuous-time weighting is not restricted to time to event outcomes; the same weights may be used for other outcomes, e.g. binary variables. The continuous-time weights are desirable when, in the case of treatment weights, the time to treatment initiation is a time to event variable, as in our prostate cancer example. Since no censoring occurred before the end of follow-up, we could have used non-survival methods e.g. weighted logistic regression to model the Failure outcome. Still, we prefer the cumulative incidences as they are easily interpreted and can be estimated without parametric assumptions.

Our analysis of Norwegian patients with localized prostate cancer suggest no medium to long term difference in failure of treatment rates among patients receiving radiation therapy and radical prostatectomy. These results are relevant for the ongoing discussion on the comparative effectiveness of RP and Rad (Wallis *and others*, 2017; Tree and Dearnaley, 2017). A recent RCT on patients with localized prostate cancer suggested that radiation therapy and radical prostatectomy gave similar rates of disease progression after 10 years of follow-up (Hamdy *and others*, 2016). However, this study has been criticized for being underpowered to detect clinically significant differences and observational studies have suggested that RP is favorable (Wallis *and others*, 2017).

We are currently working on the statistical properties of the weighting process, and methods for weighted additive hazard regressions are under development. Flexible weight estimation is implemented in the `R` package `ahw`, that will be uploaded to CRAN.

5. Supplementary material

Supplementary material, including details about the data sources, the parametric models for weight calculation, as well as comparisons between continuous-time and discrete time approaches is available at http://biostatistics.oxfordjournals.org.

6. Software

Software for estimation and assessment of continuous-time treatment weights, illustrated on simulated data examples can be found on `https://github.com/palryalen/ahw`. For questions, comments or remarks about the shared code, contact the corresponding author (p.c.ryalen@medisin.uio.no).

Acknowledgments

Conflict of Interest: None declared.

Funding

Pål Christie Ryalen, Mats Julius Stensrud, and Kjetil Røysland were supported by the research grant NFR239956/F20 - Analyzing clinical health registries: Improved software and mathematics of identifiability.

References

Aalen, O.O. (1989). A linear regression model for the analysis of life times. *Statistics in Medicine* **8**, 907–925.

Borocas, D.A., Alvarez, J., Resnick, M.J., Koyama, T., Hoffman, K.E., Tyson, M.D., Conwill, R., McCollum, D., Cooperberg, M.R, Goodman, M., Greenfield, S., Hamilton, A.S., Hashibe, M., Kaplan, S.H, Paddock, L.E., Stroup, A.M., Wu, X.C. *and others.* (2017). Association between radiation therapy, surgery, or observation for localized prostate cancer and patient-reported outcomes after 3 years. *JAMA* **317**(11), 1126–1140.

Didelez, V. (2008). Graphical models for marked point processes based on local independence. *J. R. Statist. Soc. B* **70**, 245–264.

Gray, R. (1988). A class of k-sample tests for comparing the cumulative incidence of a competing risk. *The Annals of Statistics* **16**(3), 1141–1154.

Hamdy, F.C., Donovan, J.L., Lane, J.A., Mason, M., Metcalfe, C., Holding, P., Davis, M., Peters, T.J., Turner, E.L, Martin, R.M, Oxley, J., Robinson, M., Walsh, J. Staffurthand E., Bollina, P., Catto, J., Doble, A., Doherty, A., Kockelbergh, D. Gillattand R., Kynaston, H., Paul, A., Powell, P., Prescott, S., Rosario, D.J., Rowe, E. *and others.* (2016). 10-year outcomes after monitoring, surgery, or radiotherapy for localized prostate cancer. *New England Journal of Medicine* **375**(15), 1415–1424.

Hernán, Miguel A. (2010). The hazards of hazard ratios. *Epidemiology (Cambridge, Mass.)* **21**(1), 13.

Hernán, Miguel A, Brumback, Babette A and Robins, James M. (2002). Estimating the causal effect of zidovudine on cd4 count with a marginal structural model for repeated measures. *Statistics in medicine* **21**(12), 1689–1709.

Jacod, J. (1975). Multivariate point processes: Predictable projection, radon-nikodym derivatives, representation of martingales. *Probability Theory and Related Fields* **31**, 235–253.

Joffe, Marshall M, Ten Have, Thomas R, Feldman, Harold I and Kimmel, Stephen E. (2004). Model selection, confounder control, and marginal structural models: review and new applications. *The American Statistician* **58**(4), 272–279.

Krupski, Tracey L, Kwan, Lorna, Afifi, Abdelmonem A and Litwin, Mark S. (2005). Geographic and socioeconomic variation in the treatment of prostate cancer. *Journal of clinical oncology* **23**(31), 7881–7888.

Moodie, Erica EM, Stephens, David A and Klein, Marina B. (2014). A marginal structural model for multiple-outcome survival data: assessing the impact of injection drug use on several causes of death in the canadian co-infection cohort. *Statistics in medicine* **33**(8), 1409–1425.

Pearl, J. (2000). *Causality: Models, Reasoning and Inference 2nd Edition*. Cambridge University Press.

Robins, JM. (1997). Asa proceedings of the section on bayesian statistical science. In: *American Statistical Association*. American Statistical Association. pp. 1–10.

Røysland, K. (2011). A martingale approach to continuous-time marginal structural models. *Bernoulli* **17**(3), 895–915.

Ryalen, P. C., Stensrud, M.J. and Røysland, K. (2017, October). Transforming cumulative hazard estimates. *ArXiv e-prints*.

Ryalen, P. C., Stensrud, M. J. and Røysland, K. (2018, February). The additive hazard estimator is consistent for continuous time marginal structural models. *ArXiv e-prints*.

Tree, Alison and Dearnaley, David. (2017). Randomised controlled trials remain the key to progress in localised prostate cancer. *European Urology* **73**, 21–22.

Wallis, Christopher JD, Glaser, Adam, Hu, Jim C, Huland, Hartwig, Lawrentschuk, Nathan, Moon, Daniel, Murphy, Declan G, Nguyen, Paul L, Resnick, Matthew J and Nam, Robert K. (2017). Survival and complications following surgery and radiation for localized prostate cancer: An international collaborative review. *European Urology* **73**(1), 11–20.

Woods, LM, Rachet, B and Coleman, MP. (2005). Origins of socio-economic inequalities in cancer survival: a review. *Annals of Oncology* **17**(1), 5–19.

Table 1. Summary of baseline variables($n = 1296$). In the hazard regression we coded the baseline variables Age at diagnosis, Gleason, Risk Group, PSA, and T category as in (Borocas *and others*, 2017), all binary. Education level was included as a categorical variable. The same variables were used in the logistic regression, only this time PSA was coded categorically. In addition, the logistic regression included another five Comorbidity variables.

Variable	Radiation($n = 544$)	Radical prostatectomy($n = 752$)
Age at diagnosis > 65 (%)	280 (51.5)	211 (28.1)
Gleason > 6	224 (41.2)	257 (34.2)
Risk Group 1	183 (33.6)	385 (51.2)
PSA > 5	488 (89.7)	594 (79.0)
PSA categorical		
$(0, 8]$	189 (34.7)	472 (62.8)
$(8, 15]$	256 (47.1)	248 (33.0)
$(15, 22]$	107 (19.7)	40 (5.32)
T cathegory 1	285 (52.4)	455 (60.5)
Education level		
Less than high school	135 (24.8)	156 (20.7)
More than four years college	56 (10.3)	97 (12.9)
Between the two above	353 (64.9)	499 (66.4)
Comorbidity variables(1:Yes)		
Hypertension = 1	247 (45.4)	255 (33.9)
Coronary artery disease = 1	141 (25.9)	123 (16.4)
Atrial fibrillation = 1	47 (8.64)	33 (4.39)
Diabetes = 1	37 (6.80)	23 (3.06)
Earlier cancer diagnosis = 1	22 (4.04)	27 (3.59)

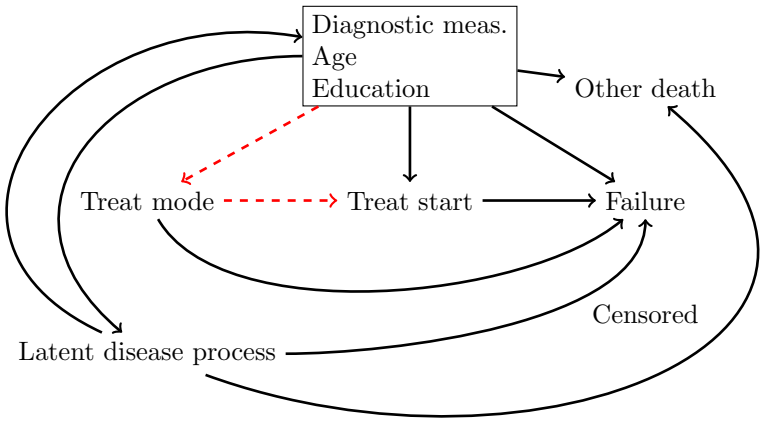

Fig. 1. Local independence graph describing the prostate cancer registry data. Treat mode denotes either radical prostatectomy (RP) or radiation therapy (RAD). Diagnostic meas. denotes factors that influence the perceived severity of the diagnosis, e.g. PSA levels and Gleason score. The dashed arrows indicate dependencies that are present in the registry data, but would not be present in our hypothetical randomized setting. Censored is included as a node, but is not associated with any arrow in the Figure for aesthetic reasons. However, an arrow from each of the other nodes should be pointing into the state, indicating the assumption that Censoring is locally independent of each and every node, conditional on the remaining ones.

[*Received September 6th, 2017; revised XX; accepted for publication YY*]

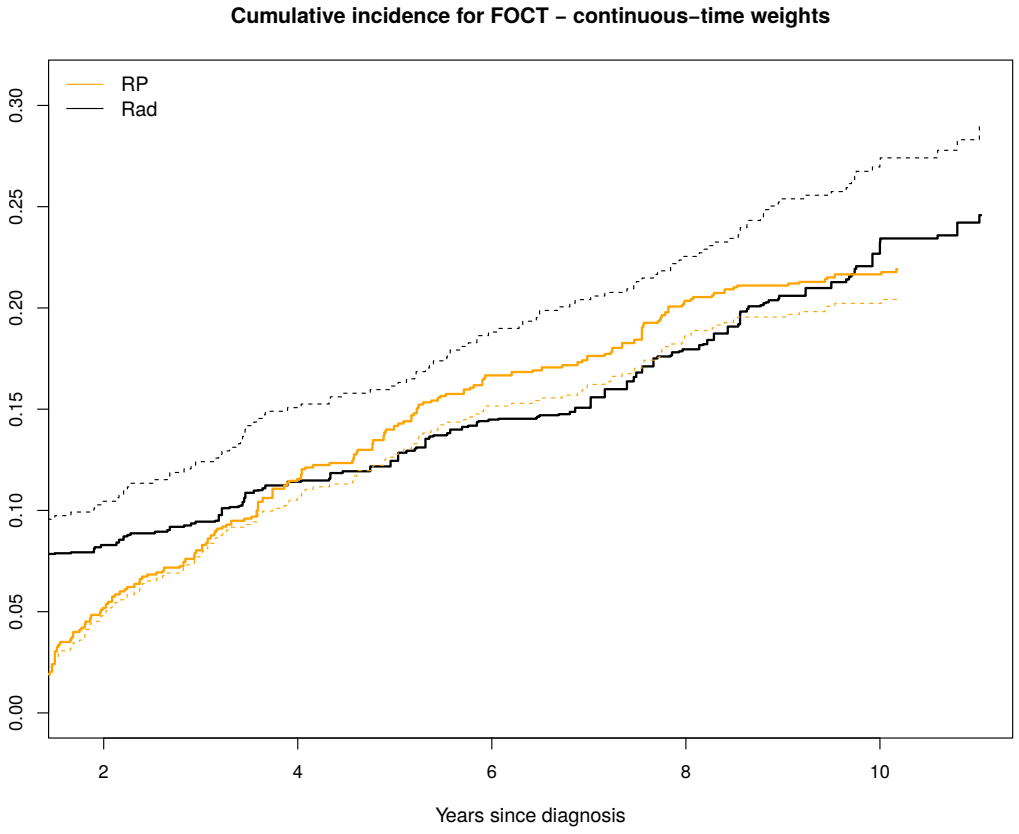

Fig. 2. Cumulative incidences for FOCT under the two treatment regimens. The weighted analysis is shown in thick solid lines, while the unweighted naive analysis is shown in dotted lines.

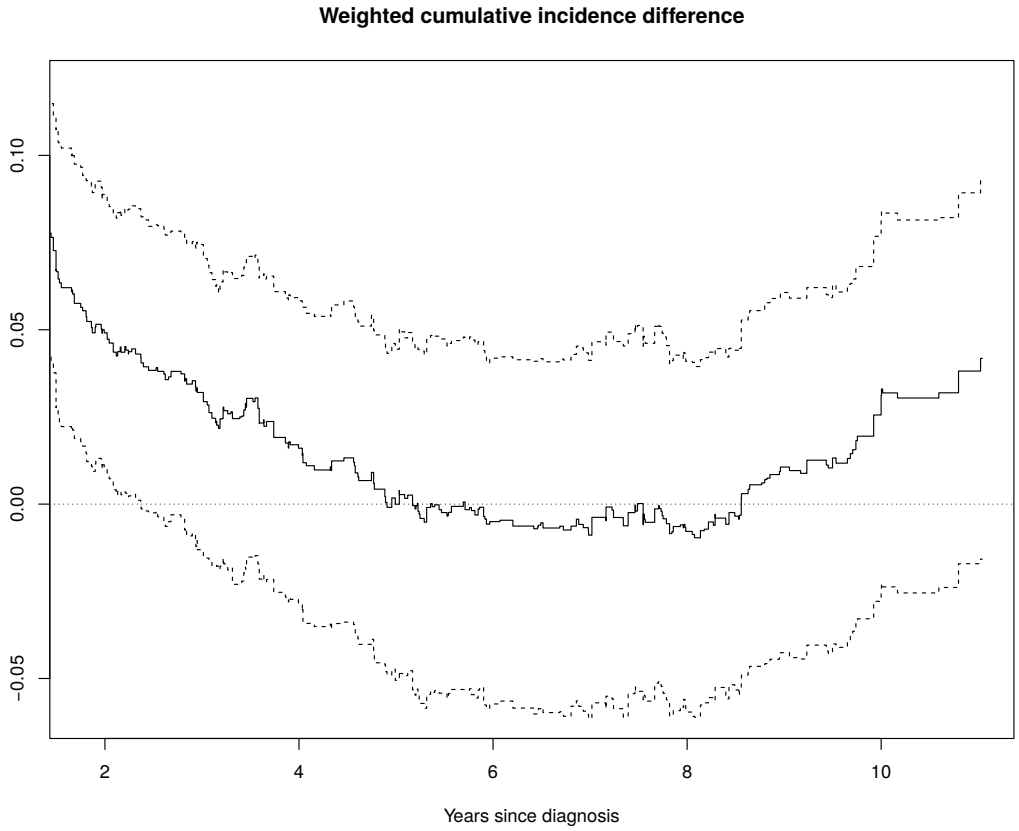

Fig. 3. Weighted cumulative incidence difference with bootstrapped 95% confidence intervals. The bootstrap sample size was 3000. By multiplying the solid curve with the sample size one can estimate the expected difference in number of treatment failures as a function of time. This estimate would be close to zero, and no more than about 60 in the period of interest(two years after diagnosis and onwards.)