Lifetime Data Analysis (2019) 25:611–638
https://doi.org/10.1007/s10985-019-09468-y

The additive hazard estimator is consistent for continuous-time marginal structural models

Pål C. Ryalen[1] · **Mats J. Stensrud**[1] · **Kjetil Røysland**[1]

Received: 7 February 2018 / Accepted: 14 February 2019 / Published online: 23 February 2019

Abstract
Marginal structural models (MSMs) allow for causal analysis of longitudinal data. The standard MSM is based on discrete time models, but the continuous-time MSM is a conceptually appealing alternative for survival analysis. In applied analyses, it is often assumed that the theoretical treatment weights are known, but these weights are usually unknown and must be estimated from the data. Here we provide a sufficient condition for continuous-time MSM to be consistent even when the weights are estimated, and we show how additive hazard models can be used to estimate such weights. Our results suggest that continuous-time weights perform better than IPTW when the underlying process is continuous. Furthermore, we may wish to transform effect estimates of hazards to other scales that are easier to interpret causally. We show that a general transformation strategy can be used on weighted cumulative hazard estimates to obtain a range of other parameters in survival analysis, and explain how this strategy can be applied on data using our `R` packages `ahw` and `transform.hazards`.

Keywords Additive hazard models · Causal inference in survival analysis · Continuous time marginal structural models · Continuous time weights

1 Introduction

MSMs can be used to obtain causal effect estimates in the presence of confounders, which e.g. may be time-dependent (Robins et al. 2000). The procedure is particularly

✉ Pål C. Ryalen
p.c.ryalen@medisin.uio.no

Mats J. Stensrud
m.j.stensrud@medisin.uio.no

Kjetil Røysland
kjetil.roysland@medisin.uio.no

[1] Department of Biostatistics, University of Oslo, Domus Medica Gaustad, Sognsvannsveien, 0372 Oslo, Norway

appealing because it allows for a sharp distinction between confounder adjustment and model selection (Joffe et al. 2004): first, we adjust for observed confounders by weighing the observed data to obtain balanced pseudopopulations. Then, we calculate effect estimates from these pseudopopulations based on our structural model.

Traditional MSM techniques for survival analysis have considered time to be a discrete processes (Hernán et al. 2000b). In particular, inverse probability of treatment weights (IPTWs) are used to create the pseudopopulations, and then e.g. several subsequent logistic regressions are fitted for discrete time intervals to mimic a proportional hazards model.

However, time is naturally perceived as a continuous process, and it also seems natural to analyse time-to-event outcomes with continuous models. Inspired by the discrete time MSMs, Røysland (2011) suggested a continuous-time analogue to MSMs. Similar to the discrete MSMs, it has been shown that the continuous MSMs can be used to obtain consistent effect estimates when the theoretical treatment weights are known (Røysland 2011). In particular, additive hazard regressions can be weighted with the theoretical continuous-time weights to yield consistent effect estimates. Nevertheless, the weights are usually unknown in real life and must be estimated from the data.

In this article, we show that continuous-time MSMs also perform desirable when the treatment weights are estimated from the data: we provide a sufficient condition to ensure that weighted additive hazard regressions are consistent. Furthermore, we show how such weighted hazard estimates can be consistently transformed to obtain other parameters that are easier to interpret causally. To do this, we use stability theory of SDEs, which allows us to target a range of parameters expressed as solutions of ordinary differential equations. Many examples of such parameters can be found in Ryalen et al. (2018b). This is immediately appealing for causal survival analysis: first, we can use hazard models, that are convenient for regression modeling, to obtain weights. Estimates on the hazard scale are hard to interpret causally per se (Robins and Greenland 1989; Hernán 2010; Aalen et al. 2015; Stensrud et al. 2017), but we present a generic method to consistently transform these effect estimates to several other scales that are easier to interpret.

The continuous-time weights and the causal parameters can be estimated using the `R` package `ahw`. We show that this `ahw` weight estimator, which is based on additive hazard regression, is consistent in Theorem 2. We have implemented code for transforming cumulative hazard estimates in the package `transform.hazards`. These packages make continuous-time marginal structural modeling easier to implement for applied researchers.

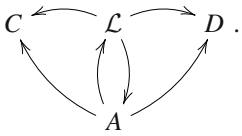
2 Weighted additive hazard regression

2.1 Motivation

We will present a strategy for dealing with confounding and dependent censoring in continuous time. Confounding, which may be time-varying, will often be a problem

when analysing observational data, e.g. coming from health registries. The underlying goal is to assess the effect a treatment strategy has on an outcome.

We can describe processes in continuous time using local (in)dependence relations, and we can use local independence graphs to visualise these relations. A precise description of local independence can be found in Røysland (2011). The local independence graph we will focus on is



Heuristically, the time-dependent confounders $\mathcal{L}$ and the exposure A can influence the censoring process C and the event of interest D. Moreover, the time-dependent confounders can both influence and be influenced by the exposure process. We include baseline variables, some of which may be confounders, in Sect. 2.2.

The above graph can e.g. describe a follow-up study of HIV-infected subjects, where the initiation and adjustment of HIV treatment depend on CD4 count measurements over time (Hernán et al. 2000a). The CD4 count is a predictor of future survival, and it is also a diagnostic factor that informs initiation of zidovudine treatment; a CD4 count below a certain threshold indicates that treatment is needed. The CD4 count will, in turn, tend to increase in response to treatment, and is monitored over time to inform the future treatment strategy. Hence, it is a time-dependent confounder. In most follow-up studies there is a possibility for subjects to be censored, and we allow the censoring to depend on the covariate and treatment history, as long as subjects are alive.

In Ryalen et al. (2018a) we analysed a cohort of Norwegian males diagnosed with prostate cancer, using the theory from this article to compare treatment effectiveness of radiation and surgery, even though time-dependent confounding were thought to be a minor issue. The continuous-time MSMs allowed us to estimate causal cumulative incidences on the desired time-scale, starting from the time of diagnosis. This example shows that (continuous-time) MSMs can also be a preferable choice in the absence of time-dependent confounding.

2.2 Hypothetical scenarios and likelihood ratios

We consider observational event-history data where n i.i.d. subjects are followed over the study period $[0, T]$. Let $N^{i,A}$ and $N^{i,D}$ respectively be counting processes that jump when treatment A and outcome D of interest occur for subject i. Furthermore, let $Y^{i,A}$, $Y^{i,D}$ be the at-risk processes for A and D. We let $\mathcal{V}_0$ be the collection of baseline variables that are not confounders, as well as the treatment and outcome processes. $\mathcal{L}$ are the (time-dependent) confounders. For now, we assume independent censoring, but we will show how our methods can be applied in some scenarios with dependent censoring in Sect. 6.

Let $\mathcal{F}_t^{i,\mathcal{V}_0\cup\mathcal{L}}$ denote the filtration that is generated by all the observable events for individual i. Moreover, let P^i denote the probability measure on $\mathcal{F}_T^{i,\mathcal{V}_0\cup\mathcal{L}}$ that governs

the frequency of observations of these events, and $\lambda_t^{i,D}$ denote the intensity for $N^{i,D}$ with respect to P^i and the filtration $\mathcal{F}_t^{i,\mathcal{V}_0\cup\mathcal{L}}$.

We aim to estimate the outcome in a hypothetical situation where a treatment intervention is made according to a specified strategy. Suppose that the frequency of observations we would have seen in this hypothetical scenario is described by another probability measure $\tilde{P}^i$ on $\mathcal{F}_T^{i,\mathcal{V}_0\cup\mathcal{L}}$. Furthermore, assume that all the individuals are also i.i.d. in the hypothetical scenario and that $\tilde{P}^i \ll P^i$, i.e. that there exists a likelihood ratio

$$R_t^i := \frac{d\tilde{P}^i|_{\mathcal{F}_t^{i,\mathcal{V}_0\cup\mathcal{L}}}}{dP^i|_{\mathcal{F}_t^{i,\mathcal{V}_0\cup\mathcal{L}}}}$$

for each time t. We will later describe how an explicit form of $\{R^i\}_i$ can be obtained. It relies on the assumption that the underlying model is causal, a concept we define in Sect. 3. For the moment we will not require this, but only assume that $\lambda_t^{i,D}$ defines the intensity with respect to $\mathcal{F}_t^{i,\mathcal{V}_0\cup\mathcal{L}}$ for both P^i and $\tilde{P}^i$; that is, the functional form of $\lambda_t^{i,D}$ is identical under both P^i and $\tilde{P}^i$.

Suppose that $N^{i,D}$ has an additive hazard with respect to $\tilde{P}^i$ and the filtration $\mathcal{F}_t^{i,\mathcal{V}_0}$ that is generated by the components of $\mathcal{V}_0$. We stress that we consider the intensity process marginalised over $\mathcal{L}$, and it is thereby defined with respect to $\mathcal{F}_t^{i,\mathcal{V}_0}$, and not $\mathcal{F}_t^{i,\mathcal{V}_0\cup\mathcal{L}}$. In other words, we assume that the hazard for event D with respect to the filtration $\mathcal{F}_t^{i,\mathcal{V}_0}$ is additive, and can be written as

$$\boldsymbol{X}_{t-}^{i\intercal}\boldsymbol{b}_t, \tag{1}$$

where $\boldsymbol{b}_t$ is a bounded and continuous vector-valued function, and the components of $\boldsymbol{X}^i$ are covariate processes or baseline variables from $\mathcal{V}_0$.

2.3 Re-weighted additive hazard regression

Our main goal is to estimate the cumulative coefficient function in (1), i.e.

$$\boldsymbol{B}_t := \int_0^t \boldsymbol{b}_s ds \tag{2}$$

from the observational data distributed according to $P = P^1 \otimes \cdots \otimes P^n$. If we had known all the true likelihood ratios, we could try to estimate (2) by re-weighting each individual in Aalen's additive hazard regression (Andersen et al. 1993, VII.4) according to its likelihood ratio. However, the true weights are unlikely to be known, even if the model is causal. In real-life situations, we can only hope to have consistent estimators for these weights. We therefore consider $\mathcal{F}_t^{1,\mathcal{V}_0\cup\mathcal{L}}\otimes\cdots\otimes\mathcal{F}_t^{n,\mathcal{V}_0\cup\mathcal{L}}$-adapted estimates $\{R_t^{(i,n)}\}_n$ that converge to R_t^i under relatively weak assumptions, such that Aalen's additive hazard regression for the outcome re-weighted according to $\{R_t^{(i,n)}\}$

gives consistent estimates of the causal cumulative hazard. The estimator we will consider is defined as follows: let $\boldsymbol{N}^{(n)}$ be the vector of counting processes and $\boldsymbol{X}^{(n)}$ the matrix containing the $\boldsymbol{X}^i$'s, that is,

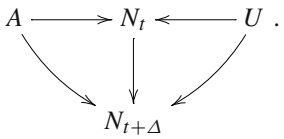
$$\boldsymbol{N}_t^{(n)} := \begin{pmatrix} N_t^{1,D} \\ \vdots \\ N_t^{n,D} \end{pmatrix} \text{ and } \boldsymbol{X}_s^{(n)} := \begin{pmatrix} X_s^{1,1} & \dots & X_s^{1,p} \\ \vdots & & \vdots \\ X_s^{n,1} & \dots & X_s^{n,p} \end{pmatrix}, \tag{3}$$

and let $\boldsymbol{Y}_s^{(n),D}$ denote the $n \times n$-dimensional diagonal matrix where the i'th diagonal element is $Y_s^{i,D} \cdot R_{s-}^{(i,n)}$. The weighted additive hazard regression is given by:

$$\boldsymbol{B}_t^{(n)} := \int_0^t (\boldsymbol{X}_{s-}^{(n)\intercal} \boldsymbol{Y}_s^{(n),D} \boldsymbol{X}_{s-}^{(n)})^{-1} \boldsymbol{X}_{s-}^{(n)\intercal} \boldsymbol{Y}_s^{(n),D} d\boldsymbol{N}_s^{(n)}. \tag{4}$$

2.3.1 Parameters that are transformations of cumulative hazards

It has recently been emphasised that the common interpretation of hazards in survival analysis as the causal risk of death during $(t, t + \Delta]$ for an individual that is alive at t, is often not appropriate; see e.g. Hernán (2010). An example in Aalen et al. (2015) shows that this can also be a problem in RCTs; if N is a counting process that jumps at the time of the event of interest, A is a randomised treatment, and U is an unobserved frailty, the following causal diagram describes such a situation:



If we consider the probability of an event before $N_{t+\Delta}$, conditioning on no event at time t, we condition on a collider that opens a non-causal path from A to the outcome. This could potentially have dramatic consequences since much of survival analysis is based on the causal interpretation of hazards, e.g. hazard ratios.

In Ryalen et al. (2018b), we have suggested a strategy to handle this situation: even if it is difficult to interpret hazard estimates causally per se, we can use hazard models to obtain other parameters that have more straightforward interpretations. Population based measures such as the survival function, the cumulative incidence functions, and the restrictive mean survival function, do not condition on survival and will therefore not be subject to the selection bias. Moreover, these measures, and many others (see Ryalen et al. 2018b; Stensrud et al. 2018 for examples), solve differential equations driven by cumulative hazards, i.e. they are functions $\boldsymbol{\eta}_t$ that can be written on the form

$$\boldsymbol{\eta}_t = \boldsymbol{\eta}_0 + \int_0^t F(\boldsymbol{\eta}_{s-}) d\boldsymbol{B}_s, \tag{5}$$

where $\boldsymbol{B}$ are cumulative hazard coefficients, and F is a Lipschitz continuous matrix-valued function. In Ryalen et al. (2018b), we showed how to estimate $\boldsymbol{\eta}$ by replacing the integrator in (5) with an estimator $\boldsymbol{B}^{(n)}$ that can be written as a counting process integral. Examples of such $\boldsymbol{B}^{(n)}$ include the Nelson–Aalen, or more generally Aalen's additive hazard estimator. This gave rise to the stochastic differential equation

$$\boldsymbol{\eta}_t^{(n)} = \boldsymbol{\eta}_0^{(n)} + \int_0^t F(\boldsymbol{\eta}_{s-}^{(n)}) d\boldsymbol{B}_s^{(n)}, \tag{6}$$

that is easy to solve on a computer; it is a piecewise constant, recursive equation that jumps whenever the integrator $\boldsymbol{B}^{(n)}$ jumps. Hence, (6) can be solved using a `for` loop over the jump times of $\boldsymbol{B}^{(n)}$, i.e. the survival times of the population.

A simple example of a parameter on the form (5) is the survival function, which reads $S_t = 1 - \int_0^t S_s dB_s$, where B is the cumulative hazard for death. In this case, the estimation strategy (6) yields the Kaplan–Meier estimator. Nevertheless, some commonly studied parameters cannot be written on the form (5), such as the median survival, and the hazard ratio.

In Ryalen et al. (2018b) we showed that $\boldsymbol{\eta}^{(n)}$ provides a consistent estimator of $\boldsymbol{\eta}$ if

- $\lim_{n\to\infty} P(\sup_{t\leq T} |\boldsymbol{B}_t^{(n)} - \boldsymbol{B}_t| \geq \epsilon) = 0$ for every $\epsilon > 0$, i.e. the cumulative hazard estimator is consistent, and
- the estimator $\boldsymbol{B}^{(n)}$ is predictably uniformly tight, abbreviated P-UT.

The additive hazard estimator satisfies both these criteria, and additive hazard regression can thus be used as an intermediate step for flexible estimation of several parameters, such as the survival, the restricted mean survival, and the cumulative incidence functions (Ryalen et al. 2018b). In Theorem 1, we show that also the reweighted additive hazard regression satisfies these properties, which is a major result in this article. Thus, we can calculate causal cumulative hazard coefficients, and transform them to estimate MSMs that solve ordinary differential equations consistently. In Sect. 4.4 we illustrate how such estimation can be done, by including an example of a marginal structural relative survival model on simulated data.

A mathematically precise definition of P-UT is given in Jacod and Shiryaev (2003, VI.6a). We will not need the full generality of this definition here. Rather, we will use (Ryalen et al. 2018b, Lemma 1) to determine if processes are P-UT. The Lemma states that whenever $\{\boldsymbol{J}_t^{(n)}\}_n$ is a sequence of semi-martingales on $[0, T]$ with Doob–Meyer decompositions

$$\boldsymbol{J}_t^{(n)} = \int_0^t \boldsymbol{\rho}_s^{(n)} ds + \boldsymbol{M}_t^{(n)},$$

where $\{\boldsymbol{M}^{(n)}\}_n$ are square integrable local martingales and $\{\boldsymbol{\rho}^{(n)}\}_n$ are predictable processes such that

$$\lim_{a\to\infty} \sup_n P\left(\sup_s |\boldsymbol{\rho}_s^{(n)}|_1 \geq a \right) = 0 \text{ and} \tag{7}$$

$$\lim_{a\to\infty} \sup_n P\left(\mathrm{Tr}\langle \boldsymbol{M}^{(n)} \rangle_T \geq a \right) = 0, \tag{8}$$

then $\{\boldsymbol{J}_t^{(n)}\}_n$ is P-UT. Here, Tr is the trace function, and $\langle \cdot \rangle$ is the predictable variation.

2.4 Consistency and P-UT property

The consistency and P-UT property of $\boldsymbol{B}^{(n)}$ introduced in Sect. 2.3 is stated as a Theorem below. A proof can be found in the "Appendix".

Theorem 1 (Consistency of weighted additive hazard regression) *Suppose that*

(I) *The conditional density of* $R_t^{(i,n)}$ *given* $\mathcal{F}_t^{i,\mathcal{V}_0 \cup \mathcal{L}}$ *does not depend on* i,

(II)

$$E_P[\sup_{t\leq T} |\lambda_t^{1,D}|^2] < \infty \text{ and } E_P[\sup_{t\leq T} |\boldsymbol{X}_t^1|^2] < \infty$$

(III) *Let*

$$\boldsymbol{\Gamma}_t^{(n)} := \left(\frac{1}{n} \boldsymbol{X}_{t-}^{(n)\intercal} \boldsymbol{Y}_t^{(n),D} \boldsymbol{X}_{t-}^{(n)} \right) = \left(\tfrac{1}{n} \textstyle\sum_{k=1}^n R_{t-}^{(k,n)} X_{t-}^{k,i} Y_t^{k,D} X_{t-}^{k,j} \right)_{i,j},$$

and suppose that

$$\lim_{a\to\infty} \inf_n P\left(\sup_{t\leq T} \mathrm{Tr}\left(\boldsymbol{\Gamma}_t^{(n)-1} \right) > a \right) = 0,$$

(IV) *Suppose that* $\{R^i\}_i$ *and* $\{R^{(i,n)}\}_{i,n}$ *are uniformly bounded and*

$$\lim_{n\longrightarrow\infty} P\left(\left| R_t^{(i,n)} - R_t^i \right| > \delta \right) = 0 \tag{9}$$

for every i, $\delta > 0$ *and* t.

Then $\{\boldsymbol{B}^{(n)}\}_n$ *is P-UT and*

$$\lim_{n\longrightarrow\infty} P\left(\sup_{t\leq T} \left| \boldsymbol{B}_t^{(n)} - \boldsymbol{B}_t \right| \geq \delta \right) = 0, \tag{10}$$

for every $\delta > 0$.

Heuristically, condition (I) states that if we know individual i's realisation of the variables and processes in $\mathcal{V}_0 \cup \mathcal{L}$ up to time t, no other information on individual i is used for estimating her weight at t. Condition (II) ensures that the number of outcome events will not blow up, or suddenly grow by an extreme amount. Condition (III) implies that there can be no collinearity among the covariates, or more precisely that the inverse matrix of $\left(E[X_t^{1,i} X_t^{1,j}] \right)_{i,j}$ is uniformly bounded in t. Condition (IV)

states that the weight estimator converges to the theoretical weights R_t^i, in a not very strong sense. The uniform boundedness of $\{R^i\}_i$ is a positivity condition similar to the positivity condition required for standard inverse probability weighting.

3 Causal validity and a consistent estimator for the individual likelihood ratios

We can model the individual likelihood ratio in many settings where the underlying model is causal. To do this, we assume that each subject is represented by the outcomes of r baseline variables $Q_1, \dots Q_r$, and d counting processes $N^1, \dots, N^d$. Moreover, we let $\mathcal{F}_t$ denote the filtration that is generated by all their possible events before t.

Suppose that $\lambda^1, \dots, \lambda^d$ are the intensities of the counting processes $N^1, \dots, N^d$ with respect to the filtration $\mathcal{F}_t$ and the observational probability P. Now, by Jacod (1975), $P|_{\mathcal{F}_T}$ is uniquely determined by all the intensities and the conditional densities at baseline of the form $dP\big(Q^k|Q^{k-1}, \dots, Q^1\big)$, because the joint density at baseline factorises as a product of conditional densities.

Suppose that the observational scenario, where the frequency of events are described by P, is subject to an intervention on the component represented by N^j. Our model is said to be **causal** if such an intervention would not change the 'local characteristics' of the remaining nodes. More precisely this means that

- The functional form of the intensities on which we do not intervene coincide under P and the intervened scenario $\tilde{P}$, i.e. λ^k would also define the intensity for N^k with respect to $\tilde{P}$ when $k \neq j$, and
- The conditional density of each Q^k, given $Q^{k-1}, \dots, Q^1$ would be the same with respect to both P and $\tilde{P}$, i.e.

$$dP\big(Q^k|Q^{k-1}, \dots, Q^1\big) = d\tilde{P}\big(Q^k|Q^{k-1}, \dots, Q^1\big)$$

for $k = 1, \cdots r$.

If the intervention instead were targeted at a baseline variable, say Q^j, and this intervention would replace $dP\big(Q^k|Q^{k-1}, \dots, Q^1\big)$ by $d\tilde{P}\big(Q^k|Q^{k-1}, \dots, Q^1\big)$, for $k = 1, \cdots r$, the model is said to be causal if

- The intensity process for N^k with respect to P and $\tilde{P}$ coincide for all $k = 1, \cdots p$, and
- The remaining conditional densities at baseline coincide, i.e.

$$dP\big(Q^k|Q^{k-1}, \dots, Q^1\big) = d\tilde{P}\big(Q^k|Q^{k-1}, \dots, Q^1\big),$$

for $k \neq j$.

Note that the latter is in agreement with Pearl's definition of a causal model (Pearl 2000).

This notion of causal validity leads to an explicit formula for the likelihood ratio. If the intervention is aimed at N^j, changing the intensity from λ^j to $\tilde{\lambda}^j$, then the likelihood ratio takes the form

$$R_t = \left(\prod_{s \leq t} \theta_s^{\Delta N_s^j} \right) \exp\left(\int_0^t \lambda_s^j - \tilde{\lambda}_s^j ds \right), \tag{11}$$

where $\theta_t := \frac{\tilde{\lambda}_t^j}{\lambda_t^j}$, see Røysland (2011) and Jacod (1975).

If the intervention is targeted at a baseline variable, the likelihood ratio corresponds to the ordinary propensity score

$$R_0 := \frac{d\tilde{P}\left(Q^j | Q^{j-1}, \ldots, Q^1\right)}{dP\left(Q^j | Q^{j-1}, \ldots, Q^1\right)}. \tag{12}$$

Interventions on several nodes yield a likelihood ratio that is a product of terms on the form (11) and (12). The terms in the product could correspond to baseline interventions, time-dependent treatment interventions, or interventions on the censoring intensity. It is natural to estimate the likelihood ratio, or weight process by a product of baseline weights, treatment weights, and censoring weights.

We want, of course, to identify the likelihood ratio that corresponds to $\tilde{P}$, as this is our strategy to assess the desired randomised trial. Following Eqs. (11) and (12), we see that the intervened intensities and baseline variables must be modeled correctly, and specifically that a sufficient set of confounders must be included when modeling the treatment intensity. Additionally, the MSM for the outcome must be correctly specified. An important consequence of the results in this paper is that a class of MSM parameters that solve ODEs driven by cumulative hazards can be estimated consistently.

As long as the intervention acts on a counting process or a baseline variable, the same formula would hold in much more general situations where the remaining covariates are represented by quite general stochastic processes. The assumption of 'coinciding intensities' must then be replaced by the assumption that the 'characteristic triples', a generalisation of intensities to more general processes, coincides for P and $\tilde{P}$; see Jacod and Shiryaev (2003, II.2).

3.1 Estimation of continuous-time weights using additive hazard regression

Suppose we have a causal model as described in the beginning of Sect. 3, allowing us to obtain a known form of the likelihood ratio R^i. To model the hypothetical scenario, we need to rely on estimates of the likelihood ratio. In the following, we will only focus on a causal model where we replace the intensity of treatment by $\tilde{\lambda}^{i,A}$, the intensity of $N^{i,A}$ with respect to P and the subfiltration $\mathcal{F}_t^{\mathcal{V}_0}$. It is a consequence of the innovation theorem (Andersen et al. 1993) that $E[\lambda_t^{i,A} | \mathcal{F}_{t-}^{\mathcal{V}_0}] = \tilde{\lambda}_t^{i,A}$. Moreover, an exercise in asymptotics of stochastic processes shows that if we discretise time, the

associated marginal model structural weights from Robins et al. (2000) approximate (11) gradually as the time-resolution increases.

We will not follow the route of Robins et al. (2000) to estimate R^i. Instead, we will use that (11) is the unique solution to the stochastic differential equation

$$\begin{aligned} R_t^i &= R_0^i + \int_0^t R_{s-}^i dK_s^i \\ K_t^i &= \int_0^t (\theta_s^i - 1) dN_s^{i,A} + \int_0^t \lambda_s^{i,A} ds - \int_0^t \tilde{\lambda}_s^{i,A} ds, \end{aligned}$$

with $\theta^i = \frac{\tilde{\lambda}^{i,A}}{\lambda^{i,A}}$. To proceed, we assume that $\lambda^{i,A}$ and $\tilde{\lambda}^{i,A}$ satisfy the additive hazard model, i.e. that there are vector valued functions $\boldsymbol{h}_t$ and $\tilde{\boldsymbol{h}}_t$, and covariate processes $\boldsymbol{Z}_t$ and $\tilde{\boldsymbol{Z}}_t$ that are adapted to $\mathcal{F}_t^{i,\mathcal{V}_0 \cup \mathcal{L}}$ and $\mathcal{F}_t^{i,\mathcal{V}_0}$ respectively, and

$$\lambda_t^{i,A} = Y_t^{i,A} \boldsymbol{Z}_t^{i\intercal} \boldsymbol{h}_t \text{ and } \tilde{\lambda}_t^{i,A} = Y_t^{i,A} \tilde{\boldsymbol{Z}}_t^{i\intercal} \tilde{\boldsymbol{h}}_t. \tag{13}$$

The previous equation translates into the following:

$$\begin{aligned} R_t^i &= R_0^i + \int_0^t R_{s-}^i dK_s^i \\ K_t^i &= \int_0^t (\theta_s^i - 1) dN_s^{A,i} + \int_0^t Y_s^{i,A} \boldsymbol{Z}_s^{i\intercal} d\boldsymbol{H}_s - \int_0^t Y_s^{i,A} \tilde{\boldsymbol{Z}}_s^{i\intercal} d\tilde{\boldsymbol{H}}_s, \end{aligned}$$

where $\boldsymbol{H}_t = \int_0^t \boldsymbol{h}_s ds$ and $\tilde{\boldsymbol{h}}_t = \int_0^t \tilde{\boldsymbol{h}}_s ds$.

Our strategy is to replace R_0^i, $\boldsymbol{H}$, $\tilde{\boldsymbol{H}}$ and θ^i by estimators. This gives the following stochastic differential equation:

$$\begin{aligned} R_t^{(i,n)} &= R_0^{(i,n)} + \int_0^t R_{s-}^{(i,n)} dK_s^{(i,n)} \\ K_t^{(i,n)} &= \int_0^t (\theta_{s-}^{(i,n)} - 1) dN_s^{i,A} + \int_0^t Y_s^{i,A} \boldsymbol{Z}_{s-}^{i\intercal} d\boldsymbol{H}_s^{(n)} - \int_0^t Y_s^{i,A} \tilde{\boldsymbol{Z}}_{s-}^{i\intercal} d\tilde{\boldsymbol{H}}_s^{(n)}, \end{aligned} \tag{14}$$

where the quantity $R_0^{(i,n)}$ is assumed to be a consistent estimator of R_0^i. We will use the additive hazard regression estimators $\boldsymbol{H}^{(n)}$ and $\tilde{\boldsymbol{H}}^{(n)}$ for estimating $\boldsymbol{H}$ and $\tilde{\boldsymbol{H}}$ (Andersen et al. 1993). Moreover, suppose that $\theta_0^{(i,n)}$ is a consistent estimator of θ_0^i, the intensity ratio evaluated at zero. Our candidate for $\theta_t^{(i,n)}$ when $t > 0$ depends on the choice of an increasing sequence $\{\kappa_n\}_n$ with $\lim_{n \longrightarrow \infty} \kappa_n = \infty$ such that $\sup_n \frac{\kappa_n}{\sqrt{n}} < \infty$. This estimator takes the form

$$\theta_t^{(i,n)} = \begin{cases} \theta_0^{(i,n)}, 0 \leq t < 1/\kappa_n \\ \frac{\int_{t-1/\kappa_n}^t Y_s^{i,A} \tilde{\boldsymbol{Z}}_{s-}^{i\intercal} d\tilde{\boldsymbol{H}}_s^{(n)}}{\int_{t-1/\kappa_n}^t Y_s^{i,A} \boldsymbol{Z}_{s-}^{i\intercal} d\boldsymbol{H}_s^{(n)}}, \quad 1/\kappa_n \leq t \leq T. \end{cases} \tag{15}$$

κ_n can thus be interpreted as a smoothing parameter. We let $\boldsymbol{Y}^{(n),A}$ be the diagonal matrix where the i'th diagonal element is $Y^{i,A}$. The following Theorem says that the above strategy works out.

Theorem 2 *Suppose that*

a. *Each θ^i is uniformly bounded, and right-continuous at $t = 0$.*
b. *For each i,*

$$\lim_{\delta \to 0} P\big(\inf_{t \leq T} |\tilde{\boldsymbol{Z}}_t^{i\intercal} \tilde{\boldsymbol{h}}_t| \leq \delta \big) = 0, \tag{16}$$

c. $E\big[\sup_{s \leq T} |\boldsymbol{Z}_s^i|_3^3 \big] < \infty$ *and* $E\big[\sup_{s \leq T} |\tilde{\boldsymbol{Z}}_s^i|_3^3 \big] < \infty$ *for every i*
d.

$$\lim_{a \to \infty} \sup_n P\left(\sup_{s \leq T} \mathrm{Tr}\left(\left(\frac{1}{n} \boldsymbol{Z}_s^{(n)\intercal} \boldsymbol{Y}_s^{(n),A} \boldsymbol{Z}_s^{(n)} \right)^{-1} \right) \geq a \right) = 0$$

and

$$\lim_{a \to \infty} \sup_n P\left(\sup_{s \leq T} \mathrm{Tr}\left(\left(\frac{1}{n} \tilde{\boldsymbol{Z}}_s^{(n)\intercal} \boldsymbol{Y}_s^{(n),A} \tilde{\boldsymbol{Z}}_s^{(n)} \right)^{-1} \right) \geq a \right) = 0$$

Then we have that

$$\lim_{n \to \infty} P\left(\sup_{t \leq T} |R_t^{(i,n)} - R_t^i| > \delta \right) = 0 \tag{17}$$

for every $\delta > 0$ and i.

For Theorem 1 to apply we need that our additive hazard weight estimator and the likelihood ratio are uniformly bounded. The latter will for instance be the case if both $\lambda^{i,A} - \tilde{\lambda}^{i,A}$ and $\tilde{\lambda}^{i,A}/\lambda^{i,A}$ are uniformly bounded. We will, however, only assume that the theoretical weights R^i are uniformly bounded. In that case we can also make our weight estimator $R^{(i,n)}$ uniformly bounded, by merely truncating trajectories that are too large.

4 Example

4.1 Software

We have developed `R` software for estimation of continuous-time MSMs that solve ordinary differential equations, in which additive hazard models are used to model both the time to treatment and the time to the outcome of interest. Our procedure involves two steps: first, we estimate continuous-time weights using fitted values of the treatment model. These weights can be used to re-weight the sample for estimating

the outcome model. Second, we take the cumulative hazard coefficients of the weighted (or causal) outcome model and transform them to estimate ODE parameters that have a more appealing interpretation than cumulative hazards. The two steps can be performed using the `R` packages `ahw` and `transform.hazards`, both of which are available in the repository `github.com/palryalen`. Below, we show an example on how to use the packages on simulated data.

4.2 A simulation study

We simulate an observational study where individuals may experience a terminating event D, so that the hazard for D depends additively on the treatment A and a covariate process L. A and L are counting processes that jump from 0 to 1 for an individual at the instant treatment is initiated or the covariate changes, respectively. The subjects receive treatment depending on L, such that L is a time-dependent confounder. The subjects in the $L = 1$ group can move into treatment, while the subjects in the $L = 0$ group may receive treatment or move to the $L = 1$ group in any order. All subjects are at risk of experiencing the terminating event. The following data generating hazards for D, A, and L are utilised:

$$\alpha_t^D = \alpha_t^{D|0} + \alpha_t^{D|A} A_{t-} + \alpha_t^{D|L} L_{t-} + \alpha_t^{D|A,L} A_{t-} L_{t-} \tag{18}$$

$$\alpha_t^A = \alpha_t^{A|0} + \alpha_t^{A|L} L_{t-}$$

$$\alpha_t^L = \alpha_t^{L|0} + \alpha_t^{L|A} A_{t-}. \tag{19}$$

We want to assess the effect of A on D we would see if A were randomised, i.e. if treatment initiation did not depend on L. To find the effect A has on D we perform a weighted analysis.

We remark that this scenario could be made more complicated by e.g. allowing the subjects to move in and out of treatment, or have recurrent treatments. We could also have included a dependent censoring process, and re-weighted to a hypothetical scenario in which censoring were randomised (see Sect. 6).

4.3 Weight calculation using additive hazard models

We assume that the longitudinal data is organised such that each individual has multiple time-ordered rows; one row for each time either A, L or D changes.

Our goal is to convert the data to a format suitable for weighted additive hazard regression. Heuristically, the additive hazard estimates are cumulative sums of least square estimations evaluated at the event times in the sample. The main function will therefore need to do two jobs; (a) the data must be expanded such that every individual, as long as he is still at risk of D, has a row for each time D occurs in the population, and (b) each of those rows must have an estimate of his weight process evaluated just before that event time.

Our software relies on the `aalen` function from the `timereg` package. We fit two additive hazard models for the transition from untreated to treated. The first model

assesses the transitions that we observe, i.e. where treatment is influenced by a subjects realisation of L. Here, we use (19), i.e. the true data generating hazard model for treatment initiation; an additive hazard model with intercept and L as a covariate. The second model describes the transitions under the hypothetical randomised trial in which each individual's treatment initiation time is a random draw of the treatment initiation times in the population as a whole. The treatment regime in our hypothetical trial is given by the marginal treatment initiation hazard of the study population, which is the hazard obtained by integrating out L from (19). We estimate the cumulative hazard using the Nelson–Aalen estimator for the time to treatment initiation, by calling a marginal `aalen` regression.

In this way we obtain a factual and a hypothetical `aalen` object that are used as inputs in our `makeContWeights` function. Other input variables include the bandwidth parameter used in (15), weight truncation options, and an option to plot the weight trajectories.

The output of the `makeContWeights` function is an expanded data frame where each individual has a row for every event time in the population, with an additional `weight` column containing time-updated weight estimates. To do a weighted additive hazard regression for the outcome, we will use the `aalen` function once again. Weighted regression is performed on the expanded data frame by setting the weights argument equal to the weight column.

When the weighted cumulative hazard estimates are at hand, we can transform our cumulative hazard estimates as suggested in Sect. 2.3.1, to obtain effect measures that are easier to interpret. This step can be performed using the `transform.hazards` package; see the GitHub vignette for several worked examples.

4.4 A marginal structural model

We now suppose the intervention that imposes a marginal treatment initiation rate is causally valid. This implies that the intensity for the event D has the same form under the randomised scenario $\tilde{P}$, i.e. that the hazard for D under $\tilde{P}$ for the filtration $\mathcal{F}_t^{A\cup D\cup L}$, generated by A, D, and L, takes the same functional form as (18). We are, however, interested in the hazard with respect to $\tilde{P}$ and the subfiltration $\mathcal{F}_t^{A\cup D}$, the filtration generated by A and D (note that $\mathcal{F}_t^{A\cup D\cup L}$ and $\mathcal{F}_t^{A\cup D}$ respectively correspond to $\mathcal{F}_t^{\mathcal{V}_0\cup\mathcal{L}}$ and $\mathcal{F}_t^{\mathcal{V}_0}$ from Sect. 2.2). By the innovation theorem the hazard with respect to $\tilde{P}$ and $\mathcal{F}_t^{A\cup D}$ takes the form

$$\beta(t|A) = \beta_t^0 + \beta_t^A A_{t-}.$$

A straightforward regression analysis of the observational data cannot yield causal estimates. Using the ideas from Sect. 2, we can estimate the cumulative coefficients $B_t^{A=0} = \int_0^t \beta_s^0 ds$ and $B_t^{A=1} - B_t^{A=0} = \int_0^t \beta_s^A ds$ consistently by performing a weighted additive hazard regression.

Cumulative hazards, however, are not easy to interpret. We therefore assess effects on the survival scale, using a marginal structural relative survival model. In this example, our marginal structural relative survival RS^A solves

$$RS_t^{A=a} = 1 + \int_0^t \left(-RS_s^{A=a} \; RS_s^{A=a}\right) d \begin{pmatrix} B_s^{A=a} \\ B_s^{A=0} \end{pmatrix}. \tag{20}$$

The quantity $RS^{A=1}$ can be understood as the survival probability a subject would have if he were exposed at time 0, relative to the survival probability he would have if he were never exposed. Our suggested plugin-estimator is obtained by inserting the estimated causal cumulative coefficients, i.e. the weighted estimates $\hat{B}^{A=a}$ and $\hat{B}^{A=0}$:

$$\hat{RS}_t^{A=a} = 1 + \int_0^t \left(-\hat{RS}_{s-}^{A=a} \; \hat{RS}_{s-}^{A=a}\right) d \begin{pmatrix} \hat{B}_s^{A=a} \\ \hat{B}_s^{A=0} \end{pmatrix}.$$

4.5 Simulation details and results

We simulate subjects, none of which are treated at baseline. Initially, all the patients start with $L = 0$, and the hazards for transitioning from one state to another is constant. As described in Sect. 4.3, we fit additive hazard models for the time to treatment initiation, one for the observed treatment scenario, i.e. (19), and one for the hypothetical randomised scenario. These models are inserted into `makeContWeights` to obtain weight estimates. Finally, we estimate the additive hazard model by calling the `aalen` function where the `weights` option is set equal to the weight column in the expanded data set.

We make comparisons to the discrete-time, stabilised IPTWs, calculated using pooled logistic regressions. To do this, we discretise the study period [0, 10] into K equidistant subintervals, and include the time intervals as categorical variables in the regressions. We fit two logistic regressions; one for the weight numerator, regressing only on the intercept and the categorical time variables, and a covariate-dependent model for the weight denominator, regressing on the intercept, the categorical time variables, and the time-updated covariate process. We then calculate IPTWs by extracting the predicted probabilities of the two logistic regression model fits, and inserting them into the cumulative product formula [Robins et al. 2000, eq. (17)].

In the upper three rows of Fig. 1 we display estimates of the causal cumulative hazard coefficient, i.e. estimates of $B^{A=1} - B^{A=0}$, for a range of sample sizes. We include estimates weighted according to our estimator (14), the IPTW estimator, and the theoretical weights, i.e. the true likelihood ratios $\{R^i\}_i$. Compared to the discrete weight estimators, our continuous-time weight estimator (14) gives better approximations to the curves that are estimated with the theoretical weights. In the lowest row of Fig. 1 we plot $\hat{RS}^{A=1}$, i.e. transformed estimates of the cumulative hazard coefficients re-weighted according to the different weight estimators. We used the `transform.hazards` package to perform the plugin-estimation.

5 Performance

In Fig. 2 we plot mean weight estimates based on aggregated simulations of the set-up in Sect. 4. The plot suggests that the discrete weights gradually approximate

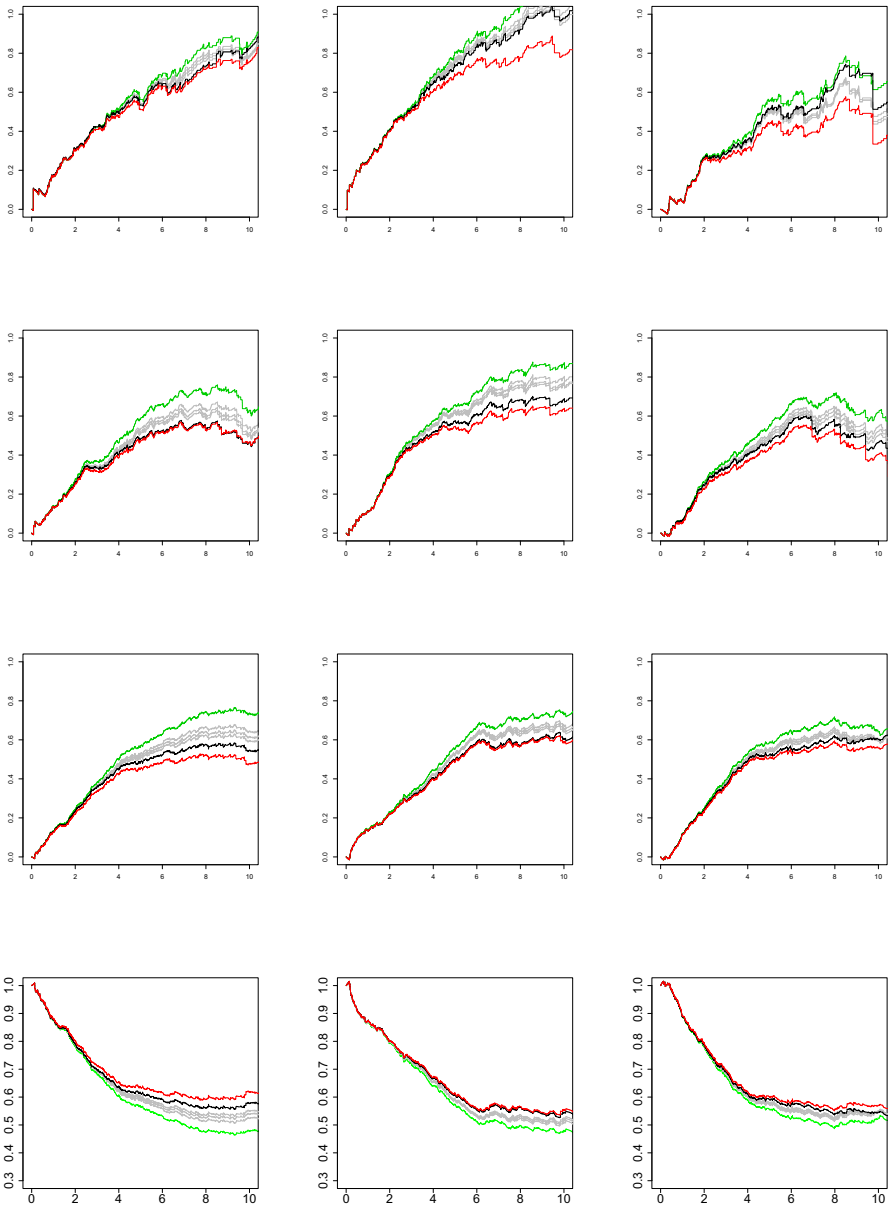


Fig. 1 The upper three rows: three realisations of the cumulative treatment effect estimates for the same scenario, with $n = 500$, 1000, and 2000 from top to bottom. A red line based on estimates re-weighted with the true R^i's is included for reference. The green line shows the unweighted estimates, the gray lines are obtained using the IPTW estimates, while the black line is obtained using our additive hazard weight estimates. The discrete weights were estimated using pooled logistic regressions based on $K = 4$, 8, and 16 time intervals. Increasing the number of intervals moved the curves closer to the red curve. The lowermost row: estimated causal effect of being treated at $t = 0$ versus never being treated according to the relative survival MSM, based on the $n = 2000$ sample (Color figure online)

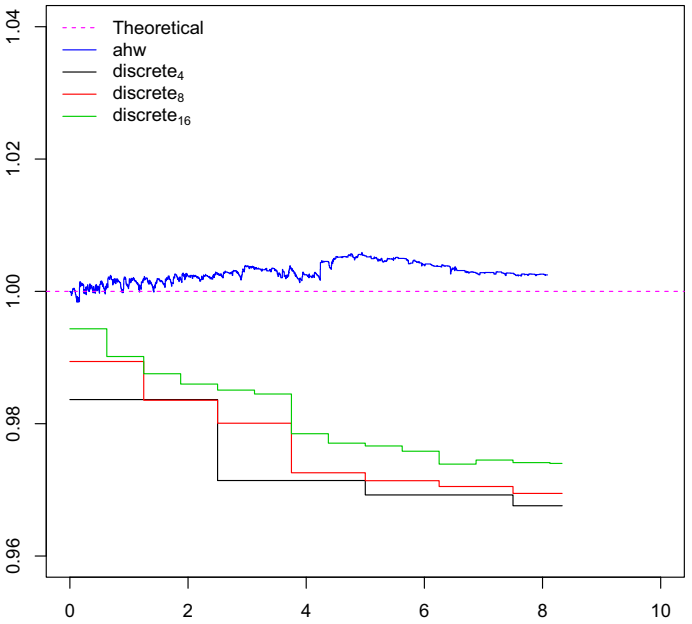


Fig. 2 Average weights based on a sample size of 3000. The theoretical weights have expected value 1. Included are our additive hazard weights, as well as IPTW with $K = 4$, 8, and 16 time intervals. We see that the discrete weights are biased approximations of the theoretical likelihood ratio, while our additive hazard weight estimator appears to be less biased

the continuous likelihood ratio as the time discretisation is refined. However, the continuous-time weights (14) are closer to the expected value of 1 at all times t, indicating less bias.

Choosing the bandwidth parameter will influence the weight estimator and weighted additive hazard estimator in a bias-variance tradeoff; a small κ_n will yield estimates with large bias and small variance, while a large κ_n will give rise to small bias but large variance. It is difficult to provide an exact recipe for choosing the bandwidth parameter, since a good choice depends on several factors, such as the sample size, the distribution of the treatment times, as well as the form and complexity of the true treatment model: if the true treatment hazard is constant, a small κ_n is often appropriate. If the treatment hazard is highly time-varying, κ_n should be chosen to be large, depending on the sample size. Heuristically, several treatment times in the interval $[t - 1/\kappa_n, t]$ for each t would be desirable, but this is not possible in every situation, e.g. when the treatment time distribution is skewed. Such distributions can lead to instable, and possibly large weights for some subjects, even if the chosen bandwidth parameter is a good choice for most other subjects. One option is to truncate weights that are larger than a specified threshold, at the cost of introducing bias. We can assess sensitivity concerning the choice of the bandwidth by performing an analysis for several bandwidth values, truncating weights if necessary, and comparing the resulting weighted estimators. This approach was taken in Ryalen et al. (2018a, see e.g. Supplementary Figure 4), where no noticeable difference was found for four values of κ_n.

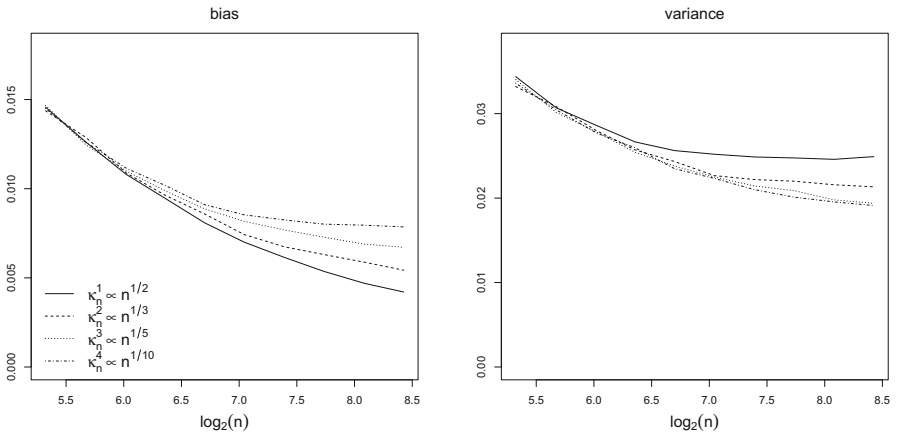


Fig. 3 Bias and variance as a function of n, for four bandwidth refinement strategies

We inspect the bias and variance of our weight estimator for sample sizes n under four bandwidth choices κ_n^z, $z = 1, 2, 3, 4$ at a specified time t_0. By aggregating estimates of k samples for each n we get precise estimates of the bias and variance as a function of n for each choice. The bandwidth functions are scaled such that they are identical at the smallest sample n_0, with $\kappa_{n_0}^1 = \kappa_{n_0}^2 = \kappa_{n_0}^3 = \kappa_{n_0}^4 = 1/t_0$. Otherwise they satisfy $\kappa_n^1 \propto n^{1/2}$, $\kappa_n^2 \propto n^{1/3}$, $\kappa_n^3 \propto n^{1/5}$, and $\kappa_n^4 \propto n^{1/10}$.

We simulate a simple scenario where time to treatment initiation depends on a binary baseline variable, such that $\lambda_t^{i,A} = Y_t^{i,A}(\alpha_t^0 + \alpha_t^A x^i)$ for individual i with at-risk indicator $Y^{i,A}$ and binary variable x^i. We calculate weights that re-weight to a scenario where the baseline variable has been marginalised out, i.e. where the treatment initiation intensity is marginal. Utilising the fact that the true likelihood ratio R^i has a constant mean equal to 1, we can find precise estimates of the bias and variance of the additive hazard weight estimator (14) at time t_0.

We plot the bias and variance of the weight estimator as a function of n under the strategies κ_n^1, κ_n^2, κ_n^3 and κ_n^4 in Fig. 3. We see that the convergence strategy κ_n^1 yields a faster relative decline in bias, but a higher variance as the sample size increases. Meanwhile, the strategy κ_n^4 has a slower decline in bias, but a smaller variance than the other strategies. Finally, the strategies κ_n^2 and κ_n^3 lie mostly between κ_n^1 and κ_n^4 both concerning bias and variance, as a function of the sample size. We also see empirical justification for the requirement $\sup_n \kappa_n / n^{1/2} < \infty$, as the variance under the strategy κ_n^1 declines very slowly as n is increased.

6 Censoring weights

Most standard martingale-based estimators in survival analysis are consistent when we have independent censoring, see Andersen et al. (1993, III.2.1). We have assumed independent censoring when conditioning on $\mathcal{V}_0$. A likely situation where this is violated is when we have independent censoring when conditioned on $\mathcal{L} \cup \mathcal{V}_0$, but have

dependent censoring if we only condition on $\mathcal{V}_0$. If the model is causal with respect to an intervention that randomises censoring sufficiently, we can model the scenario where this intervention had been applied, and censoring is independent when conditioning on $\mathcal{V}_0$. This means that many estimators that are common in survival analysis will be consistent. Suppose that $N^{i,c}$ is a counting process that jumps when individual i is censored. Moreover, let $\lambda_t^{i,c}$ denote the intensity of $N^{i,c}$ with respect to the filtration $\mathcal{F}_t^{i,\mathcal{V}_0 \cup \mathcal{L}}$, and let $\tilde{\lambda}_t^{i,c}$ denote its intensity of with respect to the filtration $\mathcal{F}_t^{i,\mathcal{V}_0}$.

Suppose that there is a meaningful intervention that would give a scenario with frequencies that are governed by $\tilde{P}$ and its intensity for censoring with respect to $\mathcal{F}_t^{i,\mathcal{V}_0 \cup \mathcal{L}}$, is replaced by $\tilde{\lambda}_t^{i,c}$. If the model is causal with respect to this intervention, the corresponding likelihood ratio process is given by

$$R_t^{i,c} = \prod_{s \leq t} \left(\frac{\tilde{\lambda}_s^{i,c}}{\lambda_s^{i,c}} \right)^{\Delta N_s^{i,c}} \exp\left(- \int_0^t \tilde{\lambda}_s^{i,c} - \lambda_s^{i,c} ds \right). \tag{21}$$

However, as we only need to apply weights to observations strictly before the time of censoring, we only need to consider

$$R_t^{i,c} = \exp\left(- \int_0^t \tilde{\lambda}_s^{i,c} - \lambda_s^{i,c} ds \right). \tag{22}$$

This process is a solution to the equation

$$R_t^{i,c} = 1 + \int_0^t R_s^{i,c} (\lambda_s^{i,c} - \tilde{\lambda}_s^{i,c}) ds. \tag{23}$$

Furthermore, we assume additive hazard models, i.e. that

$$\lambda_t^c = Y_t^{i,c} \boldsymbol{U}_{t-}^{i\intercal} \boldsymbol{g}_t \text{ and } \tilde{\lambda}_t^{i,c} = Y_t^{i,c} \tilde{\boldsymbol{U}}_{t-}^{i\intercal} \tilde{\boldsymbol{g}}_t, \tag{24}$$

for an $\mathcal{F}_t^{i,\mathcal{V}_0 \cup \mathcal{L}}$-adapted covariate process $\boldsymbol{U}^i$, and an $\mathcal{F}_t^{i,\mathcal{V}_0}$-adapted covariate process $\tilde{\boldsymbol{U}}^i$, and vector valued functions $\boldsymbol{g}$ and $\tilde{\boldsymbol{g}}$. Following Theorem 2, we see that these weights are consistently estimated by $R^{(i,n,c)}$ defined by the equation:

$$\begin{aligned} R_t^{(i,n,c)} &= 1 + \int_0^t R_{s-}^{(i,n,c)} dK_s^{(i,n,c)} \\ K_t^{(i,n,c)} &= \int_0^t Y_s^{i,c} \boldsymbol{U}_{s-}^{i\intercal} d\boldsymbol{G}_s^{(n)} - \int_0^t Y_s^{i,c} \tilde{\boldsymbol{U}}_{s-}^{i\intercal} d\tilde{\boldsymbol{G}}_s^{(n)}, \end{aligned}$$

where $\boldsymbol{G}^{(n)}$ and $\tilde{\boldsymbol{G}}^{(n)}$ are the usual additive hazards estimates of $\int_0^{\cdot} \boldsymbol{g}_s ds$ and $\int_0^{\cdot} \tilde{\boldsymbol{g}}_s ds$.

7 Discussion

Marginal structural modeling is an appealing concept for causal survival analysis. Here we have developed theory for continuous-time MSMs that may motivate the approach for practical research. Indeed, we show that the continuous-time MSMs yield consistent effect estimates, even if the treatment weights are estimated from the data. Our continuous-time weights seem to perform better than the discrete time weights when we study processes that develop in continuous time. Furthermore, our weights can be estimated using additive hazard regressions, which are easy to fit in practice. Importantly, we also show that causal effect estimates on the hazard scale, e.g. weighted cumulative hazard estimates, can be transformed consistently to estimate other parameters that are easier to interpret causally. We thereby offer a broad strategy to obtain causal effect estimates for time-to-event outcomes. Previously, Huffer and McKeague (1991) and McKeague (1987) derived results on weighted additive hazard regression, but they do not cover our needs, as our weights are estimates of likelihood ratios with respect to filtrations that are larger than the filtration for the additive hazard that we want to estimate.

Estimators of IPTWs may be unstable and inefficient, e.g. when there are strong predictors of the treatment allocation. In practice, applied researchers will often face a bias-variance tradeoff when considering confounder control and efficient weight estimation. This bias-variance tradeoff has been discussed in the literature, and weight truncation has been suggested to reduce the variance, at the cost of introducing bias; see e.g. Cole and Hernán (2008). Similar to IPTWs, and for the same reasons, our continuous-time weight estimator may be instable, and proper weight estimation requires a delicate balance between confounder control and precision in most practical situations.

We have considered the treatment process A to be a time-to-event variable, but our strategy can be generalised to handle recurrent, or piecewise constant exposures. If A is allowed to have multiple jumps, the estimation procedure becomes more complex, but the same estimators (4) and (14) can be used with few modifications. We think, however, that many important applications can be explored assuming that A is the time to an event.

A different approach that accounts for time-dependent confounding is the structural nested model, which parameterises treatment effects directly in a structural model (Robins 2014). While this procedure avoids weighting, and will often be more stable and efficient, it relies on other parametric assumptions and can be harder to implement (Vansteelandt and Sjolander 2016).

We conjecture that there is a similar consistency result as Theorem 1 when the outcome model is a weighted Cox regression. However, using a Cox model in the hypothetical scenario after marginalisation leads to restrictions on the data generating mechanisms that are not properly understood, see e.g. Havercroft and Didelez (2012). This issue is related to the non-collapsibility of the Cox model, and it is a problem regardless of the weights being used are continuous or discrete.

Funding The authors were all supported by The Research Council of Norway, Grant NFR239956/F20—Analyzing clinical health registries: Improved software and mathematics of identifiability.

Appendix: proofs

We need some lemmas to prove Theorem 1.

Lemma 1 *Suppose that* $\{V^i\}_i$ *are processes on* $[0, T]$ *such that* $\sup_i E\big[\sup_s |V_s^i|\big] < \infty$, *then*

$$\lim_{a\to\infty} \sup_n P\left(\sup_s \Big|\frac{1}{n}\sum_{i=1}^n V_s^i\Big| \geq a\right) = 0. \tag{25}$$

Proof By Markov's inequality, we have for every $a > 0$ that

$$P\left(\sup_s \Big|\frac{1}{n}\sum_{i=1}^n V_s^i\Big| \geq a\right) \leq \frac{1}{na}\sum_{i=1}^n E_P\Big[\sup_s \big|V_s^i\big|\Big],$$

which proves the claim. □

Lemma 2 (A perturbed law of large numbers) *Suppose*

(I) $p^{-1} + q^{-1} = 1$, $p < \infty$,
(II) $\{V_i\}_i \subset L^p(P)$, $\{S_i\}_i \subset L^q(P)$ *such that* $\{(V_i, S_i)\}_i$ *is i.i.d., and* V_i, S_i *are measurable with respect to a* σ*-algebra* $\mathcal{F}_i$,
(III) *Triangular array* $\{S_{(i,n)}\}_{n,i\leq n}$ *such that*

$$\lim_{n\longrightarrow\infty} P\big(|S_{(1,n)} - S_1| \geq \epsilon\big) = 0 \tag{26}$$

for every $\epsilon > 0$, *and there exists a* $\tilde{S} \in L^q(P)$ *such that* $\tilde{S} \geq |S_{(1,n)}|$ *for every* n,
(IV) *The conditional density of* $S_{(i,n)}$ *given* $\mathcal{F}_i$ *does not depend on* i.

This implies that

$$\lim_{n\longrightarrow\infty} E\left[\left|\frac{1}{n}\sum_{i=1}^n S_{(i,n)}V_i - E_P[S_1V_1]\right|\right] = 0. \tag{27}$$

Proof From the triangle inequality and condition (IV) we have that

$$\begin{aligned} E\left[\left|\frac{1}{n}\sum_{i=1}^n S_{(i,n)}V_i - \frac{1}{n}\sum_{i=1}^n S_iV_i\right|\right] &\leq \frac{1}{n}\sum_{i=1}^n E\big[\big|\big(S_{(i,n)} - S_i\big)V_i\big|\big] \\ &= E\big[\big|\big(S_{(1,n)} - S_1\big)V_1\big|\big]. \end{aligned}$$

The dominated convergence theorem implies that the last term converges to 0. Finally, the weak law of large numbers and the triangle inequality yields

$$\lim_{n\longrightarrow\infty} E\left[\left|\frac{1}{n}\sum_{i=1}^{n} S_{(i,n)}V_i - E_P[S_1V_1]\right|\right]$$
$$\leq \lim_{n\longrightarrow\infty} E\left[\left|\frac{1}{n}\sum_{i=1}^{n} S_{(i,n)}V_i - \frac{1}{n}\sum_{i=1}^{n} S_iV_i\right|\right] + E\left[\left|\frac{1}{n}\sum_{i=1}^{n} S_iV_i - E[S_1V_1]\right|\right] = 0.$$

□

Lemma 3 *$\{V_i\}_i$ i.i.d. non-negative variables in $L^2(P)$, then*

$$\lim_{n\to\infty} P\left(\frac{1}{n}\max_{i\leq n} V_i \geq \epsilon\right) = 0 \tag{28}$$

for every $\epsilon > 0$.

Proof Note that

$$P\left(\frac{1}{n}\max_{i\leq n} V_i > \epsilon\right) = 1 - P\left(\max_{i\leq n} V_i \leq \epsilon n\right) = 1 - P\left(V_1 \leq \epsilon n\right)^n$$
$$= 1 - \left(1 - P\big(V_1 > \epsilon n\big)\right)^n$$

If $n > \|V_1\|_2\epsilon^{-1}$, we therefore have by Chebyshev's inequality that

$$P\left(\frac{1}{n}\max_{i\leq n} V_i > \epsilon\right) \leq 1 - \left(1 - \frac{E[V_1^2]}{n^2\epsilon^2}\right)^n,$$

where the last term converges to 0 when $n \to \infty$ since $\lim_{n\longrightarrow\infty} n\log\left(1-\frac{E[V_1^2]}{n^2\epsilon^2}\right) = 0$ for every $\epsilon > 0$. □

Lemma 4 *Define $\gamma_s^i := Y_s^{i,D}\boldsymbol{X}_s^i\boldsymbol{b}_s$, where $\boldsymbol{X}_s^i$ is the i'th row of $\boldsymbol{X}_s^{(n)}$. If the assumptions of Theorem 1 are satisfied, then*

$$\lim_{n\longrightarrow\infty} P\left(\sup_t\left|\int_0^t \boldsymbol{\Gamma}^{(n)-1}\frac{1}{n}\sum_{i=1}^{n} R_{s-}^{(i,n)}\boldsymbol{X}_{s-}^{i\intercal}(\lambda_s^{i,D} - \gamma_s^i)ds\right| \geq \delta\right) = 0 \tag{29}$$

for every $\delta > 0$.

Proof Assumption (III) from Theorem 1 and Lemma 1 implies that

$$\lim_{J\to\infty}\inf_n P\left(\sup_t\left|\Gamma_t^{(n)-1}\frac{1}{n}\sum_{i=1}^{n} R_{t-}^{(i,n)}\boldsymbol{X}_{t-}^{i\intercal}(\lambda_t^{i,D} - \gamma_t^i)\right| > J\right) = 0. \tag{30}$$

Moreover, Lemma 2 implies that

$$\frac{1}{n}\sum_{i=1}^{n} R_{t-}^{(i,n)}\boldsymbol{X}_{t-}^{i\intercal}(\lambda_t^{i,D} - \gamma_t^i)$$

converges in probability to

$$E_P\big[R^1_{t-}\boldsymbol{X}^{1\intercal}_{t-}(\lambda^{1,D}_t - \gamma^1_t)\big]$$

However, from the innovation theorem we have that this equals

$$E_{\tilde{P}}\big[\boldsymbol{X}^{1\intercal}_{t-}(\lambda^{1,D}_t - \gamma^1_t)\big] = E_{\tilde{P}}\big[\boldsymbol{X}^{1\intercal}_{t-}(E_{\tilde{P}}[\lambda^{1,D}_t|\mathcal{F}^{1,\mathcal{V}_0}_{t-}] - \gamma^1_t)\big] = 0,$$

since $\boldsymbol{X}^1_{t-}$ and γ^1_t are $\mathcal{F}^{1,\mathcal{V}_0}_{t-}$ measurable. This and (30) enables us to apply Andersen et al. (1993, Lemma II.5.3) to obtain (29). □

Lemma 5 *Suppose that (II) and (III) from Theorem 1 are satisfied and let* $\boldsymbol{M}^{(n)}_t := \Big(N^{1,D}_t - \int_0^t \lambda^{1,D}_s ds, \ldots, N^{n,D}_t - \int_0^t \lambda^{n,D}_s ds\Big)^\intercal$. *Then*

$$\boldsymbol{\Xi}^{(n)}_t := \frac{1}{n}\int_0^t \boldsymbol{\Gamma}^{(n)-1}_s \boldsymbol{X}^{(n)\intercal}_{s-} \boldsymbol{Y}^{(n),D}_s d\boldsymbol{M}^{(n)}_s \tag{31}$$

defines a square integrable local martingale with respect to the filtration $\mathcal{F}^{1,\mathcal{V}_0\cup\mathcal{L}}_s \otimes \cdots \otimes \mathcal{F}^{n,\mathcal{V}_0\cup\mathcal{L}}_s$ *and*

$$\lim_{n\to\infty} P\bigg(\operatorname{Tr}(\langle \boldsymbol{\Xi}^{(n)}\rangle_T) \geq \delta\bigg) = 0 \tag{32}$$

for every $\delta > 0$.

Proof Writing $\boldsymbol{\lambda}^{(n)}$ for the diagonal matrix with i'th diagonal element equal to $\lambda^{i,D}$, we have that

$$\operatorname{Tr}(\langle \boldsymbol{\Xi}^{(n)}\rangle_T) = \int_0^T \frac{1}{n^2}\operatorname{Tr}\Big(\boldsymbol{\Gamma}^{(n)-1}_s \boldsymbol{X}^{(n)\intercal}_{s-} \boldsymbol{Y}^{(n),D}_s \boldsymbol{\lambda}^{(n)}_s \boldsymbol{Y}^{(n),D}_s \boldsymbol{X}^{(n)}_{s-} \boldsymbol{\Gamma}^{(n)-1}_s\Big) ds. \tag{33}$$

Moreover,

$$\begin{aligned}
&\frac{1}{n^2}\operatorname{Tr}\Big(\boldsymbol{\Gamma}^{(n)-1}_s \boldsymbol{X}^{(n)\intercal}_{s-} \boldsymbol{Y}^{(n),D}_s \boldsymbol{\lambda}^{(n)}_s \boldsymbol{Y}^{(n),D}_s \boldsymbol{X}^{(n)}_{s-} \boldsymbol{\Gamma}^{(n)-1}_s\Big) && (34)\\
&\leq \frac{1}{n^2}\operatorname{Tr}\Big(\boldsymbol{\Gamma}^{(n)-1}_s \boldsymbol{X}^{(n)\intercal}_{s-} \boldsymbol{Y}^{(n),D}_s \boldsymbol{X}^{(n)}_{s-} \boldsymbol{\Gamma}^{(n)-1}_s\Big) \max_{i\leq n} Y^{i,D}_s R^{(i,n)}_{s-} \lambda^{i,D}_s && (35)\\
&\leq \operatorname{Tr}\Big(\boldsymbol{\Gamma}^{(n)-1}_s\Big)\Big(\frac{1}{n}\max_{i\leq n}\lambda^{i,D}_s\Big)\|R^{(i,n)}\|_\infty && (36)\\
&\leq \operatorname{Tr}\Big(\boldsymbol{\Gamma}^{(n)-1}_s\Big)\Big(\frac{1}{n}\sum_{i\leq n}\lambda^{i,D}_s\Big)\|R^{(i,n)}\|_\infty && (37)
\end{aligned}$$

Now, (III), (37) and Lemma 1 implies that

$$\lim_{a\to\infty}\inf_n P\left(\sup_s \frac{1}{n^2}\operatorname{Tr}\left(\boldsymbol{\Gamma}_s^{(n)-1}\boldsymbol{X}_{s-}^{(n)\intercal}\boldsymbol{Y}_s^{(n),D}\boldsymbol{\lambda}_s^{(n)}\boldsymbol{Y}_s^{(n),D}\boldsymbol{X}_{s-}^{(n)}\boldsymbol{\Gamma}_s^{(n)-1}\right)\geq a\right)=0.$$

On the other hand, Lemma 3, (36) and (III) gives us that

$$\lim_{n\to\infty} P\left(\frac{1}{n^2}\operatorname{Tr}\left(\boldsymbol{\Gamma}_s^{(n)-1}\boldsymbol{X}_{s-}^{(n)\intercal}\boldsymbol{Y}_s^{(n),D}\boldsymbol{\lambda}_s^{(n)}\boldsymbol{Y}_s^{(n),D}\boldsymbol{X}_{s-}^{(n)}\boldsymbol{\Gamma}_s^{(n)-1}\right)\geq \delta\right)=0$$

for every s and $\delta > 0$, so Andersen et al. (1993, Propositon II.5.3) implies that (31) also holds. □

Proof of Theorem 1 We have the following decomposition:

$$\begin{aligned}\boldsymbol{B}_t^{(n)}-\boldsymbol{B}_t &= \int_0^t (\boldsymbol{X}_{s-}^{(n)\intercal}\boldsymbol{Y}_s^{(n),D}\boldsymbol{X}_{s-}^{(n)})^{-1}\big(\boldsymbol{X}_{s-}^{(n)\intercal}\boldsymbol{Y}_s^{(n),D}\boldsymbol{\lambda}_s^{(n)}-\boldsymbol{X}_{s-}^{(n)\intercal}\boldsymbol{Y}_s^{(n),D}\boldsymbol{X}_{s-}^{(n)}b_s\big)ds\\ &\quad+\int_0^t (\boldsymbol{X}_{s-}^{(n)\intercal}\boldsymbol{Y}_s^{(n),D}\boldsymbol{X}_{s-}^{(n)})^{-1}\boldsymbol{X}_{s-}^{(n)\intercal}\boldsymbol{Y}_s^{(n),D}d\boldsymbol{M}_s^{(n)}\\ &= \int_0^t \boldsymbol{\Gamma}^{(n)-1}\frac{1}{n}\sum_{i=1}^n R_{s-}^{(i,n)}\boldsymbol{X}_{s-}^{i\intercal}(\lambda_s^{i,D}-\gamma_s^i)ds+\boldsymbol{\Xi}_t^{(n)}.\end{aligned}$$

Leglarts inequality (Jacod and Shiryaev 2003, Lemma I.3.30) together with Lemma 5 implies that $\boldsymbol{\Xi}^{(n)}$ converges uniformly in probability to 0. Moreover, Lemma 4 implies that $\int_0^\cdot \boldsymbol{\Gamma}^{(n)-1}\frac{1}{n}\sum_{i=1}^n R_{s-}^{(i,n)}\boldsymbol{X}_{s-}^{i\intercal}(\lambda_s^{i,D}-\gamma_s^i)ds$ converges in same sense to 0, which proves the consistency.

To see that $\boldsymbol{B}^{(n)}$ is P-UT, note that it coincides with the sum of $\boldsymbol{B}_t$, $\boldsymbol{\Xi}^{(n)}$ and $\int_0^\cdot \boldsymbol{\Gamma}_s^{(n)-1}\frac{1}{n}\sum_{i=1}^n R_{s-}^{(i,n)}\boldsymbol{X}_{s-}^{i\intercal}(\lambda_s^i-\gamma_s^i)ds$. According to Ryalen et al. (2018b, Lemma 1), the latter is P-UT since (III) and Lemma 1 implies (7). Moreover, $\boldsymbol{B}_t=\int_0^\cdot \boldsymbol{b}_s ds$ is clearly P-UT, since $\boldsymbol{b}_t$ is uniformly bounded. $\boldsymbol{\Xi}^{(n)}$ is also P-UT since Lemma 5 implies that (8) is satisfied. Finally, as $\boldsymbol{B}^{(n)}$ is a sum of three processes that are P-UT, it is necessarily P-UT itself. □

Proof of Theorem 2

Lemma 6 *Suppose that c. and d. from Theorem 2 are satisfied, and that*

(I)

$$\lim_{a\to\infty}\sup_n P\left(\sup_t\left|\theta_t^{(i,n)}\right|\geq a\right)=0,$$

(II) $\theta_{t-}^{(i,n)}$ *converges to* θ_t^i *in probability for each* i *and* t.

Then we have that $K^{(i,n)}$ is predictably uniformly tight (P-UT) and

$$\lim_n P\left(\sup_t \left|K_t^{(i,n)} - K_t^i\right| \geq \epsilon\right) = 0 \tag{38}$$

for every i and $\epsilon > 0$.

Proof Note that

$$\begin{aligned} K_t^{(i,n)} - K_t^i = &\int_0^t (\theta_{s-}^{(i,n)} - \theta_s^i) dN_s^{i,A} + n^{-1/2} \int_0^t Y_s^i \boldsymbol{Z}_{s-}^{i\intercal} d\boldsymbol{W}_s^{(n)} \\ &- n^{-1/2} \int_0^t Y_s^{i,A} \tilde{\boldsymbol{Z}}_{s-}^{i\intercal} d\tilde{\boldsymbol{W}}_s^{(n)}, \end{aligned} \tag{39}$$

where $\boldsymbol{W}_t^{(n)} := n^{1/2}(\boldsymbol{H}_t^{(n)} - \boldsymbol{H}_t)$ and $\tilde{\boldsymbol{W}}_t^{(n)} := n^{1/2}(\tilde{\boldsymbol{H}}_t^{(n)} - \tilde{\boldsymbol{H}}_t)$ are square-integrable martingales with respect to $\mathcal{F}_t^{1,\mathcal{V}_0 \cup \mathcal{L}} \otimes \cdots \otimes \mathcal{F}_t^{n,\mathcal{V}_0 \cup \mathcal{L}}$ and $\mathcal{F}_t^{1,\mathcal{V}_0} \otimes \cdots \otimes \mathcal{F}_t^{n,\mathcal{V}_0}$ respectively.

Let τ be an optional stopping time and note that

$$\begin{aligned} E\left[\left|\int_0^\tau (\theta_{s-}^{(i,n)} - \theta_s^i) dN_s^{i,A}\right|\right] &\leq E\left[\int_0^\tau \left|\theta_{s-}^{(i,n)} - \theta_s^i\right| dN_s^{i,A}\right] \\ = E\left[\int_0^\tau \left|\theta_{s-}^{(i,n)} - \theta_s^i\right| \lambda_s^{i,A} ds\right], \end{aligned}$$

so by Lenglarts inequality, (Jacod and Shiryaev 2003, I.3.30), we see that

$$\lim_{n \longrightarrow \infty} P\left(\sup_{t \leq T} \left|\int_0^t (\theta_{s-}^{(i,n)} - \theta_s^i) dN_s^{i,A}\right| \geq \epsilon\right) = 0 \tag{40}$$

for every $\epsilon > 0$ if

$$\lim_{n \longrightarrow \infty} P\left(\int_0^T \left|\theta_{s-}^{(i,n)} - \theta_s^i\right| \lambda_s^{i,A} ds \geq \epsilon\right) = 0, \tag{41}$$

for every $\epsilon > 0$. The latter property holds due to (I), (II) and Andersen et al. (1993, Proposition II.5.3).

Since $\{\int_0^t Y_s^{i,A} \boldsymbol{Z}_{s-}^{i\intercal} d\boldsymbol{W}_s^{(n)}\}_n$ converges in the skorokhod topology, we have that $\{\sup_{t \leq T} |\int_0^t Y_s^{i,A} \boldsymbol{Z}_{s-}^{i\intercal} d\boldsymbol{W}_s^{(n)}|\}_n$ is tight (Jacod and Shiryaev 2003, Theorem VI.3.21). Therefore, we also get that

$$\lim_{n \longrightarrow \infty} P\left(\sup_{t \leq T} |n^{-1/2} \int_0^t Y_s^{i,A} \boldsymbol{Z}_{s-}^{i\intercal} d\boldsymbol{W}_s^{(n)}| \geq \epsilon\right) = 0 \tag{42}$$

for every $\epsilon > 0$. For the same reason we also have

$$\lim_{n\longrightarrow\infty} P\left(\sup_{t\leq T} |n^{-1/2}\int_0^t Y_s^{i,A}\tilde{\boldsymbol{Z}}_{s-}^{i\intercal} d\tilde{\boldsymbol{W}}_s^{(n)}| \geq \epsilon\right) = 0. \tag{43}$$

By combining (42), (43) and (40), we obtain that

$$\lim_{n\longrightarrow\infty} P\left(\sup_{t\leq T} |K_t^{(i,n)} - K_t^i| \geq \epsilon\right) = 0 \tag{44}$$

for every $\epsilon > 0$.

To see that $K^{(i,n)}$ is P-UT, note that the compensator of $\int_0^{\cdot}(\theta_{s-}^{(i,n)} - 1)dN_s^{i,A}$ equals $\int_0^{\cdot}(\theta_{s-}^{(i,n)} - 1)\lambda_s^{i,A}ds$ and

$$\left\langle \int_0^{\cdot}(\theta_{s-}^{(i,n)} - 1)dN_s^{i,A} - \int_0^{\cdot}(\theta_{s-}^{(i,n)} - 1)\lambda_s^{i,A}ds \right\rangle_T = \int_0^T (\theta_{s-}^{(i,n)} - 1)^2\lambda_s^{i,A}ds.$$

The assumptions (I) in this Lemma and c) together with Ryalen et al. (2018b, Lemma 1) therefore imply that $\int_0^{\cdot}(\theta_{s-}^{(i,n)} - 1)dN_s^{i,A}$ is P-UT.

To see that $\int_0^{\cdot} Y_s^i\tilde{\boldsymbol{Z}}_{s-}^{i\intercal} d\tilde{\boldsymbol{H}}_s^{(n)}$ is P-UT, note that

$$\int_0^{\cdot} Y_s^i\tilde{\boldsymbol{Z}}_{s-}^{i\intercal} d\tilde{\boldsymbol{H}}_s^{(n)} = n^{-1/2}\int_0^{\cdot} Y_s^i\tilde{\boldsymbol{Z}}_{s-}^{i\intercal} d\tilde{\boldsymbol{W}}_s^{(n)} + \int_0^{\cdot} Y_s^i\tilde{\boldsymbol{Z}}_{s-}^{i\intercal} d\tilde{\boldsymbol{H}}_s. \tag{45}$$

An analogous decompositon yields that $\int_0^{\cdot} Y_s^i\boldsymbol{Z}_{s-}^{i\intercal} d\boldsymbol{H}_s^{(n)}$ is P-UT. This means that $K^{(i,n)}$ is a sum of three processes that are P-UT, and must therefore be P-UT itself. □

Lemma 7 *Suppose that*

(I) $\{\kappa_n\}_n$ *increasing sequence of positive numbers such that*

$$\lim_{n\longrightarrow\infty} \kappa_n = \infty \text{ and } \sup_n \frac{\kappa_n}{\sqrt{n}} < \infty,$$

(II) $\boldsymbol{h}_t$ *is a bounded and continuous vector valued function,*
(III) $\boldsymbol{Z}^i$ *is caglad with* $E[\sup_{t\leq T} |\boldsymbol{Z}_t^i|_3^3] < \infty$,
(IV)

$$\lim_{J\to\infty}\sup_n P\left(\text{Tr}\left((\frac{1}{n}\boldsymbol{Z}_{t-}^{(n)\intercal}\boldsymbol{Y}_t^{(n),A}\boldsymbol{Z}_{t-}^{(n)})^{-1}\right) \geq J\right) = 0 \tag{46}$$

(V) $Y^{i,A}\boldsymbol{Z}_{\cdot-}^{i\intercal}\boldsymbol{h}$ *defines the intensity for* $N^{i,A}$ *with respect to* P *and* $\mathcal{F}_{\cdot}^{i,\mathcal{V}_0}$. *Now,*

$$\lim_{n\longrightarrow\infty} P\left(\sup_{1/\kappa_n\leq t\leq T}\left|\kappa_n\int_{t-1/\kappa_n}^t Y_s^{i,A}\boldsymbol{Z}_{s-}^{i\intercal} d\boldsymbol{H}_s^{(n)} - Y_t^{i,A}\boldsymbol{Z}_{t-}^{i\intercal}\boldsymbol{h}_t\right| \geq \epsilon\right) = 0. \tag{47}$$

Proof Note that

$$\kappa_n \int_{t-1/\kappa_n}^{t} Y_s^{i,A} \boldsymbol{Z}_{s-}^{i\intercal} d\boldsymbol{H}_s^{(n)} - Y_t^{i,A} \boldsymbol{Z}_{t-}^{i\intercal} \boldsymbol{h}_t \tag{48}$$

$$= \frac{\kappa_n}{\sqrt{n}} \int_0^t Y_s^{i,A} \boldsymbol{Z}_{s-}^{i\intercal} d\boldsymbol{W}_s^{(n)} - \frac{\kappa_n}{\sqrt{n}} \int_0^{t-1/\kappa_n} Y_s^{i,A} \boldsymbol{Z}_{s-}^{i\intercal} d\boldsymbol{W}_s^{(n)} \tag{49}$$

$$+ \kappa_n \int_{t-1/\kappa_n}^{t} Y_s^{i,A} \boldsymbol{Z}_{s-}^{i\intercal} \boldsymbol{h}_s ds - Y_t^{i,A} \boldsymbol{Z}_{t-}^{i\intercal} \boldsymbol{h}_t. \tag{50}$$

The martingale central limit theorem implies that $\{\boldsymbol{W}^{(n)}\}$ is a sequence of martingales that converges in law to a continuous Gaussian processes with independent increments, see Andersen et al. (1993). Moreover, Ryalen et al. (2018b, Proposition 1) says that $\{\boldsymbol{W}^{(n)}\}_n$ is P-UT.

Therefore Jacod and Shiryaev (2003, Theorem VI 6.22) implies that $\int_0^{\cdot} Y_s^{i,A} \boldsymbol{Z}_{s-}^{i\intercal} d\boldsymbol{W}_s^{(n)}$ converges in law to a continuous process, so it is C-tight. Moreover, from Jacod and Shiryaev (2003, Proposition VI.3.26) we have that

$$\lim_{n\longrightarrow\infty} P\left(\sup_{1/\kappa_n \le t \le T} \left| \int_0^t Y_s^{i,A} \boldsymbol{Z}_{s-}^{i\intercal} d\boldsymbol{W}_s^{(n)} - \int_0^{t-1/\kappa_n} Y_s^{i,A} \boldsymbol{Z}_{s-}^{i\intercal} d\boldsymbol{W}_s^{(n)} \right| \ge \epsilon \right) = 0 \tag{51}$$

for every $\epsilon > 0$. The mean value theorem of elementary calculus implies that

$$\lim_{n\longrightarrow\infty} \sup_{1/\kappa_n \le t \le T} \left| \kappa_n \int_{t-1/\kappa_n}^{t} Y_s^{i,A} \boldsymbol{Z}_{s-}^{i\intercal} \boldsymbol{h}_s ds - Y_t^{i,A} \boldsymbol{Z}_{t-}^{i\intercal} \boldsymbol{h}_t \right| = 0 \tag{52}$$

P a.s. Combining (51) and (52) yields the claim. □

Proof of Theorem 2 Combining (16) and the decomposition in the proof of Lemma 7, we see that

$$\lim_{n\longrightarrow\infty} P\left(\sup_{1/\kappa_n \le t \le T} \left| \kappa_n \int_{t-1/\kappa_n}^{t} Y_s^{i,A} \tilde{\boldsymbol{Z}}_{s-}^{i\intercal} d\tilde{\boldsymbol{H}}_s^{(n)} / \tilde{\lambda}_t^{i,A} - 1 \right| \ge \epsilon \right) = 0. \tag{53}$$

Combining (16) and a. we also have

$$\lim_{n\longrightarrow\infty} P\left(\sup_{1/\kappa_n \le t \le T} \left| \kappa_n \int_{t-1/\kappa_n}^{t} Y_s^{i,A} \boldsymbol{Z}_{s-}^{i\intercal} d\boldsymbol{H}_s^{(n)} / \lambda_t^{i,A} - 1 \right| \ge \epsilon \right) = 0. \tag{54}$$

Whenever $t \geq 1/\kappa_n$, we have that by the continuous mapping theorem that

$$\begin{aligned}
&\lim_{n\longrightarrow\infty} P\left(\sup_{1/\kappa_n \leq t \leq T} \left|\theta_t^{(i,n)} - \theta_t^i\right| \geq \epsilon\right) \\
&= \lim_{n\longrightarrow\infty} P\left(\sup_{1/\kappa_n \leq t \leq T} \left|\theta_t^i \left(\frac{\kappa_n \int_{t-1/\kappa_n}^t Y_s^{i,A} \tilde{\boldsymbol{Z}}_{s-}^{i\intercal} d\tilde{\boldsymbol{H}}_s^{(n)} / \tilde{\lambda}_t^{i,A}}{\kappa_n \int_{t-1/\kappa_n}^t Y_s^{i,A} \boldsymbol{Z}_{s-}^{i\intercal} d\boldsymbol{H}_s^{(n)} / \lambda_t^{i,A}} - 1\right)\right| \geq \epsilon\right) \\
&= 0.
\end{aligned}$$

Since θ^i is right-continuous at $t = 0$, we have that

$$\lim_{n\longrightarrow\infty} P\left(\sup_{0 \leq t \leq T} \left|\theta_t^{(i,n)} - \theta_t^i\right| \geq \epsilon\right) = 0. \tag{55}$$

Finally, Jacod and Shiryaev (2003, Corollary VI 3.33) implies that $\{(R_0^{(i,n)}, K^{(i,n)})\}_n$ converges to (R_0^i, K^i) in probability. Since $K^{(i,n)}$ is P-UT,

$$R_t^{(i,n)} = 1 + \int_0^t R_{s-}^{(i,n)} dK_s^{(i,n)}$$

and

$$R_t^i = 1 + \int_0^t R_{s-}^i dK_s^i$$

Jacod and Shiryaev (2003, Theorem IX 6.9) implies that $R^{(i,n)}$ converges to R^i in probability. □

References

Aalen O, Cook R, Røysland K (2015) Does cox analysis of a randomized survival study yield a causal treatment effect? Lifetime Data Anal 21(4):579–593

Andersen P, Borgan Ø, Gill R, Keiding N (1993) Statistical models based on counting processes. Springer series in statistics. Springer, New York. ISBN 0-387-97872-0

Cole S, Hernán M (2008) Constructing inverse probability weights for marginal structural models. Am J Epidemiol 168(6):656–664

Havercroft W, Didelez V (2012) Simulating from marginal structural models with time-dependent confounding. Stat Med 31(30):4190–4206

Hernán M (2010) The hazards of hazard ratios. Epidemiology 21(1):13

Hernán M, Brumback B, Robins J (2000a) Marginal structural models to estimate the causal effect of zidovudine on the survival of HIV-positive men. Epidemiology 11(5):561–570 ISSN 10443983. http://www.jstor.org/stable/3703998

Hernán M, Brumback B, Robins J (2000b) Marginal structural models to estimate the causal effect of zidovudine on the survival of hiv-positive men. Epidemiology 11(5):561–570

Huffer F, McKeague I (1991) Weighted least squares estimation for Aalen's additive risk model. J Am Stat Assoc 86(413):114–129 ISSN 01621459. http://www.jstor.org/stable/2289721

Jacod J (1975) Multivariate point processes: predictable projection, Radon–Nikodym derivatives, representation of martingales. Probab Theory Relat Fields 31:235–253

Jacod J, Shiryaev A (2003) Limit theorems for stochastic processes. In: Grundlehren der Mathematischen Wissenschaften [Fundamental principles of mathematical sciences], vol 288, 2nd edn. Springer, Berlin, ISBN 3-540-43932-3

Joffe M, Ten Have T, Feldman H, Kimmel S (2004) Model selection, confounder control, and marginal structural models: review and new applications. Am Stat 58(4):272–279

McKeague I (1987) Asymptotic theory for weighted least squares estimators in Aalen's additive risk model

Pearl J (2000) Causality: models, reasoning and inference, 2nd edn. Cambridge University Press, Cambridge

Robins J (2014) Structural nested failure time models. Wiley StatsRef: Statistics Reference Online, New York

Robins J, Greenland S (1989) The probability of causation under a stochastic model for individual risk. Biometrics 45(4):1125–1138

Robins J, Hernán M, Brumback B (2000) Marginal structural models and causal inference in epidemiology. Epidemiology 11(5):550–560

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

Ryalen P, Stensrud M, Fosså S, Røysland K (2018a) Causal inference in continuous time: an example on prostate cancer therapy. Biostatistics. https://doi.org/10.1093/biostatistics/kxy036

Ryalen P, Stensrud M, Røysland K (2018b) Transforming cumulative hazard estimates. Biometrika. https://doi.org/10.1093/biomet/asy035

Stensrud M, Valberg M, Røysland K, Aalen O (2017) Exploring selection bias by causal frailty models: the magnitude matters. Epidemiology 28(3):379–386

Stensrud M, Røysland K, Ryalen P (2018) On null hypotheses in survival analysis. ArXiv e-prints, July

Vansteelandt S, Sjolander A (2016) Revisiting g-estimation of the effect of a time-varying exposure subject to time-varying confounding. Epidemiol Methods 5(1):37–56