The Annals of Statistics
2017, Vol. 45, No. 2, 461–499
DOI: 10.1214/15-AOS1433
© Institute of Mathematical Statistics, 2017

MIMICKING COUNTERFACTUAL OUTCOMES TO ESTIMATE CAUSAL EFFECTS

BY JUDITH J. LOK[1]

Harvard School of Public Health

In observational studies, treatment may be adapted to covariates at several times without a fixed protocol, in continuous time. Treatment influences covariates, which influence treatment, which influences covariates and so on. Then even time-dependent Cox-models cannot be used to estimate the net treatment effect. Structural nested models have been applied in this setting. Structural nested models are based on counterfactuals: the outcome a person would have had had treatment been withheld after a certain time. Previous work on continuous-time structural nested models assumes that counterfactuals depend deterministically on observed data, while conjecturing that this assumption can be relaxed. This article proves that one can mimic counterfactuals by constructing random variables, solutions to a differential equation, that have the same distribution as the counterfactuals, even given past observed data. These "mimicking" variables can be used to estimate the parameters of structural nested models without assuming the treatment effect to be deterministic.

1. Introduction. Observational studies are no replacement for randomized clinical trials, but they can be used, for example, where randomization is unethical or to generate hypotheses for subsequent clinical trials. In an observational study, treatment may be adapted to patient characteristics which predict the outcome of interest. This is called confounding by indication. If the confounding by indication only takes place at baseline, one can condition on initial person characteristics in order to get meaningful estimates of the treatment effect. However, if the confounding by indication also takes place after baseline, variables used for treatment decisions may be influenced by past treatment. Thus, they may themselves be indications of the treatment effect, and in that case simply conditioning on them can lead to false conclusions.

With such time-dependent confounding by indication, even the time-dependent Cox model does not estimate the net effect of treatment (see, e.g., [17, 20] or [21]).

Received November 2013; revised December 2015.

[1]This work was sponsored by the Netherlands Organization for Scientific Research (NWO) with a Talent scholarship, and by the National Institutes of Health, NIAID R01AI100762. The content is solely the responsibility of the author and does not necessarily represent the official views of the National Institutes of Health.

MSC2010 subject classifications. Primary 62P10; secondary 62M99, 62N02.

Key words and phrases. Causality in continuous time, dynamic treatments, longitudinal data, observational studies, panel data, rank preservation, stochastic differential equations, structural nested models.

With a time-dependent Cox model, the rate of events given past treatment and covariate history can be estimated, but the true parameter(s) on treatment may not reflect the treatment effect. A consistent estimator of the effect of the treatment on the outcome of interest has to take into account the effect of treatment on intermediate covariates. This is easily understood when considering a treatment which affects the outcome only because it affects an intermediate variable L. In that situation, if L and treatment are both included in the time-dependent Cox model for the event of interest, the true parameter(s) on treatment in this Cox model equal 0. However, treatment could be beneficial due to its effect on L. On the other hand, not including L may also result in an inconsistent estimator, if L predicts future treatment. This follows from the same reasoning as why, in case of nonrandomized point treatment, one needs to adjust for predictors of both the treatment and the outcome to consistently estimate the treatment effect: if one does not adjust for L, and if persons with L indicating a bad prognosis are more likely to be treated, the treatment may seem to adversely affect the outcome, even if it has no effect on anyone. To conclude, with time-dependent confounding by indication, one needs to take confounders into account, but adding the confounders to an outcome model is not enough.

If all confounders are measured (see Assumption 5 below), structural nested models, proposed in [18, 19, 22], and marginal structural models, proposed in [6, 26], can be used to consistently estimate treatment effects in the presence of time-dependent confounding by indication. Structural nested models and marginal structural models make a distinction between the effect of the treatment and the reason why the treatment was given, by separately modeling the treatment decisions and the treatment effect. Robins [23] compares structural nested models and marginal structural models. The current article focuses on structural nested models.

Structural nested models model relations between counterfactual outcomes. We allow for general treatment regimes. Consider a single person, who received a particular treatment regime with outcome Y. For example, the particular treatment regime could be as follows: first, no treatment, then after a certain time, initiation of treatment, then the dosage changed some time thereafter, then treatment stopped, initiated again, et cetera. Had the person's treatment been stopped (prematurely) at time t and not been re-initiated thereafter (or, had treatment changed to a "baseline" treatment regime $\overline{0}$ from time t onwards), the person's outcome, $Y^{(t)}$, might have been different.

Structural nested distribution models compare "treatment as given until time t and then changed to no treatment," leading to the outcome $Y^{(t)}$, with "treatment as given until time $t+h$ and then changed to no treatment," leading to the outcome $Y^{(t+h)}$, for $h > 0$; or, in the case where $\overline{0}$ is taken to be a "baseline treatment regime" different from "no treatment," structural nested distribution models compare "treatment as given until time t and then changed to the baseline treatment regime," leading to $Y^{(t)}$, with "treatment as given until time $t+h$ and then changed

to the baseline treatment regime," leading to $Y^{(t+h)}$. "Treatment as given" is the treatment strategy used in the population from which the data are a random sample, a treatment strategy which depends on the decisions of doctors and patients and which is generally not the same for all patients. "Treatment as given" does not need to be "no treatment" followed by "treatment initiated and not discontinued" or "treatment initiated and then permanently stopped." In addition, $\overline{0}$ can be any "baseline treatment regime," like "no treatment" or "treatment initiated and not discontinued," but which can also include starts, stops and dosage changes. In the applications described in [8, 9, 15, 24], $\overline{0}$ was "no treatment," and it was assumed that once treatment was initiated, it was never stopped. This article follows the more general approach of [22]: when treatment effects are estimated, structural nested distribution models compare "treatment as given" with the "baseline treatment regime," $\overline{0}$.

Since $Y^{(t)}$ is generally not observable, it is a counterfactual outcome. In a discrete-time setting, [5] show that existence of counterfactuals places no restrictions on the distribution of the observed variables. No comparable proof exists for the continuous-time case.

An important controversy in the causal literature is that counterfactuals are often assumed to depend deterministically on the observed data: given the model and the parameter values, all counterfactual outcomes $Y^{(t)}$ for each person can simply be calculated from the observed data. Robins [22] calls this local rank preservation (in most cases, this implies global rank preservation), when the counterfactual outcomes are solutions to the differential equation (9) in Section 4 below. Treatment is then said *not* to affect the outcome of interest if the outcome for any particular person would have been exactly the same regardless of which treatment was given. The assumption of deterministic dependence is related to the assumption of constant treatment effect in [7]: that is, the difference between counterfactual outcomes belonging to different treatments is a constant identical for all persons.

The assumption of deterministic dependence/ (local) rank preservation has frequently been attacked. This assumption does not hold if, for example, two persons with the same observed data (e.g., both receiving some prophylactic drug) could have had a different outcome had they not been treated starting from some time t (e.g., one might have contacted a virus and the other might not). In addition, deterministic dependence can never be tested, with only one outcome observed for each person. For these two reasons, the assumption of (local) rank preservation should be avoided if at all possible.

In discrete time, when treatment and covariates change at fixed times which are the same for all persons, the theory of structural nested models is well developed. Lok et al. [14] prove that it is not necessary to assume a deterministic treatment effect. In order to do so, they show that a certain "blipped down" outcome $X(t)$ mimics the outcome $Y^{(t)}$ had treatment been withheld from time t onwards, in the sense that $X(t)$ has the same distribution as $Y^{(t)}$ given past treatment and covariate

history. They also indicate why the resulting estimators for treatment effect are consistent and asymptotically normal.

However, in reality covariates and treatment often change in continuous time. Moreover, in discrete time the interpretation of the treatment effect (shift- or blip function) depends on the time scale chosen. In continuous time, the treatment effect (infinitesimal shift function) can often be interpreted as speed or rate. For these reasons, [8, 9, 15, 19, 24] have applied continuous-time structural nested distribution models. However, because of a lack of theory for these models, the applications have relied on the assumption of (local) rank preservation. The models fitted in [8, 9, 24] are described in Examples 3.1 and 3.2. Section 5 or, in greater detail, [13] describes how to use the results in the current article in order to show that assuming (local) rank preservation is not necessary to estimate treatment effects with structural nested models (an example can be found in Section 9). Therefore, the main contribution of the current article is to show that the methods in [8, 9, 15, 19, 22, 24] are robust to violation of the assumption of (local) rank preservation.

Structural nested models in continuous time are meant to estimate the effect of a continuous treatment, for which the effect of a small duration is small. Robins [22] conjectures that the appealing large sample properties of discrete-time structural nested models extend to continuous time; however, his proof requires the assumption of (local) rank preservation. He conjectures that (i) also without (local) rank preservation, a certain "blipped down" outcome $X(t)$ has the same distribution as $Y^{(t)}$ given past treatment and covariate history, (ii) the resulting estimators are consistent and asymptotically normal, and (iii) for certain models, estimators and confidence intervals can be calculated with standard software, used in a nonstandard way. This article proves conjecture (i), which we call mimicking counterfactual outcomes, and explains why such a subtle result is true. Lok [13] proves conjecture (ii), using conjecture (i). Lok [12] proves conjecture (iii), using a partial likelihood approach and conjectures (i) and (ii). Thus, the current article fills the final link in this methodology to estimate treatment effects of time-varying treatments in longitudinal observational studies without relying on (local) rank preservation. This methodology can be applied to longitudinal observational data, to study the effects of interventions affecting, for example, economic and health outcomes.

This article is organized as follows. Section 2 introduces the setting and notation of this article. Section 3 introduces the model for treatment effect, and shows some examples. Section 4 defines the mimicking variables $X(t)$ as the solution to a differential equation with a final condition. Section 4 also states the main result of this article: $X(t)$ mimics $Y^{(t)}$ in the sense that it has the same distribution as $Y^{(t)}$, even given past treatment and covariate history. Section 5 formalizes the assumption of no unmeasured confounding, which as shown there is needed to use the result of the current article to estimate the treatment effect. Section 5 also indicates how, using the mimicking result, tests and estimators can be developed without assuming (local) rank preservation. Section 6 outlines the proof of the main result of this article: $X(t)$ mimics $Y^{(t)}$ in the sense that it has the same distribution as $Y^{(t)}$, even

given past treatment and covariate history. Section 7 proves the main result of this article for nonsurvival outcomes Y. Section 8 proves the main result for survival time outcomes Y. Section 9 describes a simulation study. Section 10 concludes this article with a discussion.

2. Setting and notation. The setting to which continuous-time structural nested models apply is as follows. The outcome of interest is a continuous real-valued variable Y. For example, Y is a person's survival time, time to clinical AIDS, the number of white blood cells, or the CD4 count. Our objective is to estimate the effect of treatment on Y. In this article, we consider a fixed time interval $t \in [0, \tau]$ with finite τ, where $t = 0$ is the time at which follow-up of interest starts (e.g., 0 could be the time of enrollment in a study, or a baseline time). During the time interval $[0, \tau]$, treatment and person characteristics are observed for each person. Y is measured at or after time τ, or, in the case of a survival time outcome, Y could be measured before time τ if the person dies before time τ. We assume that treatment starts at or after time 0. We suppose that after time τ, treatment is stopped or switched to some kind of baseline treatment regime. Most of this article assumes that there is no censoring, and Y is observed for every person in the study. Section 8.4 incorporates right censoring.

The covariate process describes the course of the disease of a person, for example, the course of the blood pressure and the white blood cell count. The covariates which *must* be included are those which both (i) influence a doctor's treatment decisions *and* (ii) predict a person's prognosis with respect to the outcome of interest. If such covariates are not observed the assumption of no unmeasured confounding (see Section 5) will not hold.

Denote the probability space by $(\Omega, \mathcal{F}, P)$. For the moment, consider a single person. Write $Z(t)$ for the covariate- *and* treatment values at time t. This article assumes that $Z(t)$ takes values in $\mathbb{R}^m$, and that $Z(t) : \Omega \to \mathbb{R}^m$ is measurable for each $t \in [0, \tau]$. Moreover, we assume that Z, seen as a function on $[0, \tau]$, is continuous from the right with limits from the left (cadlag), and that with probability one this function, or "sample path," has only finitely many jumps. We also assume that the probability that the covariate and treatment process Z jumps at time t equals 0 for every fixed time t (except possibly for finitely many fixed times t, which could have point masses). For example, the hazard of the jumps of the treatment process could be continuous for all t, and could follow a continuous parametric distribution. $\overline{Z}_t = (Z(s) : 0 \le s \le t)$ denotes the covariate and treatment history until time t, and $\overline{\mathcal{Z}}_t$ is the space of cadlag functions from $[0, t]$ to $\mathbb{R}^m$ in which $\overline{Z}_t$ takes its values. Similarly, $\overline{Z}$ denotes the complete covariate and treatment history of the person in the interval $[0, \tau]$, and $\overline{\mathcal{Z}}$ is the space in which $\overline{Z}$ takes its values. In this article, the σ-algebra on $\overline{\mathcal{Z}}_t$ and $\overline{\mathcal{Z}}$ is the projection σ-algebra; measurability of $Z(s)$ for each $s \le t$ is then equivalent to measurability of the random variable $\overline{Z}_t$.

Counterfactual outcomes were already mentioned in the Introduction. $Y^{(t)}$ is the final outcome had treatment been stopped (prematurely) at time t and not been re-initiated thereafter (or changed to some kind of baseline treatment regime $\overline{0}$ from time t onwards). This article supposes that all counterfactual outcomes $Y^{(t)}$, for $t \in [0, \tau]$ and for each person, are random variables on the probability space $(\Omega, \mathcal{F}, P)$. We assume that observations and counterfactual outcomes of different persons are independent and identically distributed, and are a random sample from a larger infinite population of interest. For notational convenience, we suppress the subscript i for person.

3. Model for treatment effect. Structural nested models in continuous time model distributional relations between $Y^{(t)}$ and $Y^{(t+h)}$, for $h > 0$ small, through a so-called infinitesimal shift-function D. Write F for the cumulative distribution function and $F^{-1} : (0, 1) \mapsto \mathbb{R}$ for its generalized inverse $F^{-1}(p) = \inf\{x : F(x) \geq p\}$. Then the infinitesimal shift-function D is defined as

$$D(y, t; \overline{Z}_t) = \frac{\partial}{\partial h}\bigg|_{h=0} (F^{-1}_{Y^{(t+h)}|\overline{Z}_t} \circ F_{Y^{(t)}|\overline{Z}_t})(y), \tag{1}$$

the right-hand derivative of the quantile-quantile transform which moves quantiles of the distribution of $Y^{(t)}$ to quantiles of the distribution of $Y^{(t+h)}$ ($h \geq 0$), given the covariate and treatment history until time t, $\overline{Z}_t$. In order to define D, no assumptions are necessary about the joint distribution of the counterfactuals $Y^{(t)}$.

EXAMPLE 3.1 (Survival of AIDS patients). [24] describe an AIDS clinical trial to study the effect of AZT treatment on survival in HIV-positive patients. Time 0 was the time of enrollment in the study. Embedded within this trial was an uncontrolled observational study of the effect of prophylaxis therapy for PCP on survival. Pneumocystis Carinii Pneumonia (PCP) is an opportunistic infection that affects HIV-positive patients. Robins et al. [24] use continuous-time structural nested models to study the effect of PCP prophylaxis therapy on survival of HIV-positive patients. Thus, the outcome of interest, Y, is the survival time, and the treatment under study is prophylaxis for PCP. Although [24] estimate the effect of changes in the time the treatment is discontinued, we will consider estimating the effect of changes in the initiation time of the treatment. This conforms better to the clinical practice in HIV/AIDS, where PCP prophylaxis is rarely discontinued, and to the assumption in [24] that once PCP prophylaxis is started, it is never stopped. Therefore, in this example, $Y^{(t)}$ is the counterfactual outcome had PCP prophylaxis treatment been as given in reality until time t, initiated or continued at time t, and continued thereafter. Thus, the baseline treatment regime $\overline{0}$ in this example is "continuously treat with PCP prophylaxis." In the context of this example, the local rank preservation assumption of [24] can be expressed as

$$Y^{(t)} - t = \int_t^Y e^{\psi 1_{\text{no prophylaxis at } s}}\, ds. \tag{2}$$

Assumption (2) is very strong, because it requires that given the model parameter ψ and the observed outcome Y, all counterfactual outcomes $Y^{(t)}$ can be calculated from the observed data. The current article proves that it suffices to assume that

$$D_\psi(y,t;\overline{Z}_t) = \bigl(1-e^{\psi}\bigr)1_{\{\text{no prophylaxis at } t\}}. \tag{3}$$

This article shows that under Assumption (3),

$$Y^{(t)} - t \sim \int_t^Y e^{\psi 1_{\text{no prophylaxis at } s}}\,ds \tag{4}$$

conditional on $\overline{Z}_t$ and $Y > t$, where $\sim$ means "has the same distribution as." Given $\overline{Z}_t$, both $Y^{(t)} - t$ and $\int_t^Y e^{\psi 1_{\text{no prophylaxis at } s}}\,ds$ are random variables, depending on $Y^{(t)}$ and Y, respectively. Assumption (3) does not impose that $Y^{(t)} - t$ is equal to $\int_t^Y e^{\psi 1_{\text{no prophylaxis at } s}}\,ds$, but only that the distribution of these two random variables is the same conditional on $\overline{Z}_t$ and $Y > t$. Thus, under equation (3), patients who have the exact same observed history over $[0,\tau]$, $\overline{Z}_\tau$ and Y, do not necessarily have the same counterfactual outcomes $Y^{(t)}$. This is a substantial relaxation of the assumptions previously adopted in the literature on continuous-time structural nested models. Relaxing assumption (2) is empirically relevant because in clinical practice $Y^{(t)}$ may differ between two patients with the exact same observed history. Suppose, for example, that two patients with the exact same observed history were both on PCP prophylaxis. If one of the patients got in contact with pneumococcal bacteria (and, therefore, might have caught PCP without the preventive treatment, PCP prophylaxis), and the other did not get in contact with pneumococcal bacteria (and, therefore, might not have caught PCP, even without PCP prophylaxis), the outcomes for the two patients without PCP prophylaxis could be different.

In equation (4), the part of the residual survival time, $Y - t$, that is untreated gets multiplied by e^{ψ} to attain the same distribution as $Y^{(t)} - t$ (the residual survival time under "continuous treatment from t onwards"), conditional on $\overline{Z}_t$ and $Y > t$. Therefore, analogous to accelerated failure time models (see, e.g., Cox and Oakes [3]), the multiplication factor e^{ψ} can be interpreted in a distributional way.

Our results do not depend on adopting the particular specification of the infinitesimal shift-function D of equation (3). For example, they also apply to an alternative specification of D from [24]. In this alternative specification, the effect of the PCP prophylaxis can depend on the AZT treatment the patient received and whether or not the patient had a history of PCP prior to the start of PCP prophylaxis. Because the data in [24] were from a clinical trial for AZT treatment, AZT treatment is described by a single variable R indicating the treatment arm the patient was randomized to (R equals 1 or 2). Let $P(t)$ be equal to 1 if the patient had PCP before or at time t *and* before prophylaxis treatment started; otherwise $P(t)$ is equal to 0. The model described in [24], but adapted to our choice of baseline treatment regime ($\overline{0}$ is continuous treatment with PCP prophylaxis), is

$$D_{\psi_1,\psi_2,\psi_3}(y,t;\overline{Z}_t) = \bigl(1-e^{\psi_1+\psi_2 P(t)+\psi_3 R}\bigr)1_{\{\text{no prophylaxis at } t\}}. \tag{5}$$

This article shows that if equation (5) holds, then

$$Y^{(t)} - t \sim \int_t^Y e^{1_{\{\text{no prophylaxis at } s\}}(\psi_1+\psi_2 P(s)+\psi_3 R)}\, ds \qquad \text{given } \overline{Z}_t, \tag{6}$$

for $t < Y$.

EXAMPLE 3.2 (Effect of Graft versus Host Disease (GvHD) on time to leukemic relapse). Keiding [8] and [9] use continuous-time structural nested models to study the effect of GvHD on time to leukemic relapse in patients who had Bone Marrow Transplantation (BMT). Infection with Cytomegalovirus (CMV) is a time-dependent confounder: an independent prognostic factor for relapse that both 1. predicts the subsequent development of the exposure GvHD and 2. is predicted by past exposure GvHD. Write Y for the time until leukemic relapse. Assume that Y is observed for every patient. In [8] and [9], $Y^{(t)}$ is the outcome had the patient been exposed (or not) to GvHD as in reality until time t, and not exposed afterwards. Based on biological knowledge, [8] and [9] assume that

$$D_\psi(y, t; \overline{Z}_t) = \left(1 - e^{\psi}\right) 1_{\{\text{GvHD at } t\}}. \tag{7}$$

This article shows that then, for $t < Y$, preventing GvHD from t onwards leads to

$$Y^{(t)} - t \sim \int_t^Y e^{\psi 1_{\{\text{GvHD at } s\}}}\, ds \qquad \text{given } \overline{Z}_t. \tag{8}$$

Keiding [8] and [9] assume that (8) is true even with $\sim$ replaced by $=$ (although only for $t = 0$), hoping that assumption could be relaxed. This article shows that indeed (7) is sufficient to estimate the effect of GvHD.

EXAMPLE 3.3 (Incorporating a-priori biological knowledge, following [22]). Again consider survival as the outcome of interest. Suppose that it is known that treatment received at time t only affects survival for patients who would die by time $t + 5$ if they would receive no further treatment. An example would be a setting in which failure is death from an infectious disease, the treatment is a preventive antibiotic treatment which is of no benefit unless the person is already infected and, if death occurs, it always does within five weeks from the time of initial unrecorded subclinical infection. In that case, the natural restriction on D is that

$$D(y, t; \overline{Z}_t) = 0 \qquad \text{if } y - t > 5.$$

As can be seen from these examples, the parameters of a continuous-time structural nested model are often rates. More biostatistical examples of models for D can be found in, for example, [15, 19, 22, 25, 27].

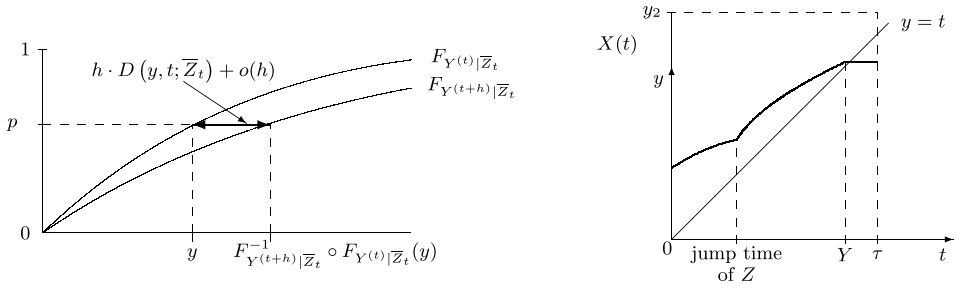

FIG. 1. Left: *Illustration of the infinitesimal shift-function* D. Right: *An example of a solution* $X(t)$ *to the differential equation* $dX(t)/dt = D(X(t), t; \overline{Z}_t)$ *with final condition* $X(\tau) = Y$ *in case the outcome is survival time.*

$h \cdot D(y, t; \overline{Z}_t)$ can be interpreted as the infinitesimal effect on the outcome of the treatment actually given in the time interval $[t, t+h)$ (relative to the baseline treatment regime). To be more precise, from the definition of D, it follows that

$$h \cdot D(y, t; \overline{Z}_t) = \big(F^{-1}_{Y^{(t+h)}|\overline{Z}_t} \circ F_{Y^{(t)}|\overline{Z}_t}\big)(y) - y + o(h).$$

In Figure 1 (left), this is sketched.

It can be shown that $D \equiv 0$ if and only if treatment does not affect the outcome of interest, as was conjectured in [22]. To be more precise, [11] shows that, for example, $D \equiv 0$ if and only if for every $h > 0$ and t, $Y^{(t+h)}$ has the same distribution as $Y^{(t)}$ given $\overline{Z}_t$. That is, $D \equiv 0$ if and only if "at any time t, whatever person characteristics are selected at that time ($\overline{Z}_t$), switching 'treatment as given' to 'baseline treatment regime' at some fixed time after t would not change the distribution of the outcome in persons with these person characteristics." To prove this, one needs the mimicking result of the current article.

4. Mimicking counterfactual outcomes. Define $X(t)$ as the continuous solution to the differential equation:

$$dX(t)/dt = D\big(X(t), t; \overline{Z}_t\big) \tag{9}$$

with final condition $X(\tau) = Y$, the observed outcome (see Figure 1, right). Then $X(t)$ mimics $Y^{(t)}$ in the sense that it has the same distribution as $Y^{(t)}$, even given the person's treatment and covariate history at time t, $\overline{Z}_t$. To prove this main result, we need the following consistency assumption.

ASSUMPTION 4.1 (Consistency). $Y^{(\tau)}$ has the same distribution as Y given $\overline{Z}_\tau$.

Notice that because no treatment was given after time τ and the treatment process is right continuous, there is no difference in treatment between $Y^{(\tau)}$ and Y. Under this consistency assumption and regularity conditions only, it is proved in

Sections 7 and 8 that indeed (9) has a unique solution X for every $\omega \in \Omega$, and that this solution $X(t)$ mimics $Y^{(t)}$ in the sense that it has the same distribution as $Y^{(t)}$ given $\overline{Z}_t$:

THEOREM 4.2 (Mimicking counterfactual outcomes). *Suppose that regularity conditions 7.1–7.4 from Section 7.2 are satisfied. Then $D(y, t; \overline{Z}_t)$ exists. Furthermore, for every $\omega \in \Omega$ there exists exactly one continuous solution $X(t)$ to $dX(t)/dt = D(X(t), t; \overline{Z}_t)$ with final condition $X(\tau) = Y$. If also consistency Assumption 4.1 is satisfied, then this $X(t)$ has the same distribution as $Y^{(t)}$ given $\overline{Z}_t$ for all $t \in [0, \tau]$.*

EXAMPLE 4.3 (Survival of AIDS patients (continuation of Example 3.1)). If equation (3) holds, then

$$X(t) = t + \int_t^Y e^{\psi 1_{\{\text{prophylaxis at } s\}}}\, ds$$

for $t < Y$, and $X(t) = Y$ for $t \geq Y$. Alternatively, if equation (5) holds, then

$$X(t) = t + \int_t^Y e^{1_{\{\text{prophylaxis at } s\}}(\psi_1 + \psi_2 P(s) + \psi_3 R)}\, ds$$

for $t < Y$, and $X(t) = Y$ for $t \geq Y$.

5. Estimators, tests, and "no unmeasured confounding". This section contains a brief summary of [13], who shows how the result of the current article leads to testing and estimation. In addition, Lok [10] Section C provides an example of estimation in our simulation study.

The main assumption underlying structural nested models is that all information the doctors used to make treatment decisions, and which is predictive of the person's prognosis with respect to the final outcome, is available for analysis. This assumption of no unmeasured confounding makes it possible to distinguish between treatment effect and selection bias; see, for example, [14, 22, 24] or [13].

Assume that the treatment process gives rise to a counting process $N(t)$. For example, $N(t)$ is the number of treatment changes until time t. The assumption of no unmeasured confounding is then formalized as the following.

ASSUMPTION 5.1 (No unmeasured confounding). The rate with which N jumps given $\overline{Z}_{t-}$ is the same as the rate with which $N(t)$ jumps given $\overline{Z}_{t-}$ and $(Y^{(s)} : s < t)$.

Because given the observed $\overline{Z}_{t-}$, the (unobserved) prognosis of a person, represented by $Y^{(s)}$ for $s < t$, should not predict treatment at or after time t. If it does, there is no way to distinguish between the effect of the treatment and the reason why it is initiated.

Notice that if $X(t)$ mimics $Y^{(t)}$ in the sense that it has the same distribution as $Y^{(t)}$ given $\overline{Z}_{t-}$, it can be expected that under no unmeasured confounding, the rate with which $N(t)$ jumps at time t also does not depend on $X(t)$, given $\overline{Z}_{t-}$. It can formally be shown that this is indeed true.

First, consider how this leads to testing. If treatment does not affect the outcome of interest, $D \equiv 0$ and thus $X(t) \equiv Y$. So if treatment does not affect the outcome of interest, changes of treatment at time t should be independent of Y, given $\overline{Z}_{t-}$. Thus, one can test whether treatment affects the outcome of interest by testing whether, given $\overline{Z}_{t-}$, Y adds to the prediction model for treatment changes.

Also, for estimation of the infinitesimal shift-function D, we assume that there is no unmeasured confounding. Suppose that one has a correctly specified parametric model D_ψ for D. Then one can calculate "$X_\psi(t)$," the solution to

$$dX_\psi(t)/dt = D_\psi\big(X_\psi(t), t; \overline{Z}_t\big) \tag{10}$$

with final condition $X(\tau) = Y$. If $X(t)$ mimics $Y^{(t)}$, then $X_\psi(t)$ has the same distribution as $Y^{(t)}$ given $\overline{Z}_t$ for the true ψ. Since $Y^{(t)}$ does not add to the prediction model for treatment changes given $\overline{Z}_{t-}$, ψ could then be estimated by picking the ψ for which, given $\overline{Z}_{t-}$, $X_\psi(t)$ adds the least to the prediction model for N, treatment changes. This can be proven to lead to the following theorem.

THEOREM 5.2. *Suppose that the intensity process λ is bounded, $Y^{()}$ is cadlag, there is no unmeasured confounding and no instantaneous treatment effect* [*with probability* 1, $N()$ *and* $Y^{()}$ *do not jump at the same time*]. *Suppose also that for every* $t \in [0, \tau]$, $X(t)$ *has the same distribution as* $Y^{(t)}$ *given* $\overline{Z}_t$. *Then*

$$E\int_0^\tau h_t\big(X(t), \overline{Z}_{t-}\big)\big(dN(t) - \lambda(t)\,dt\big) = 0$$

for each h_t *satisfying a regularity restriction. Thus, if* D_ψ *and* λ_θ *are correctly specified parametric models for* D *and* λ, *respectively*,

$$P_n\int_0^\tau h_t\big(X_\psi(t), \overline{Z}_{t-}\big)\big(dN(t) - \lambda_\theta(t)\,dt\big) = 0,$$

with P_n *the empirical measure* $P_nX = 1/n\sum_{i=1}^n X_i$, *is an unbiased estimating equation for* (θ_0, ψ_0), *for each* h_t *satisfying a regularity restriction.* h_t *here is allowed to depend on* ψ *and* θ, *as long as it satisfies the regularity restriction for* (θ_0, ψ_0).

In fact, the estimating equations of Theorem 5.2 are often martingales at the true parameter.

There is an extensive literature on the asymptotic behavior of estimators that solve unbiased estimating equations of the form $P_nU(\psi) = 0$, with $EU(\psi) = 0$; see, for example, [28] Chapter 5 for an overview. Under regularity conditions, estimators that solve unbiased estimating equations of the form $P_nU(\psi) = 0$ are

consistent and asymptotically normal. Theorem 5.2, the main result of [13], provides such unbiased estimating equations for the parameter of interest ψ, with $U(\psi)$ equal to $\int_0^\tau h_t(X_\psi(t), \overline{Z}_{t-})(dN(t) - \lambda_\theta(t)\,dt)$, provided that the main result of the current article, Theorem 4.2, holds. If θ is unknown, it can be estimated using standard methods, and the estimating equations for θ could be stacked with the estimating equations for ψ to estimate both θ and ψ. The theory from [28] Chapter 5 could be applied to those stacked estimating equations, implying that, under regularity conditions, the resulting estimator for ψ is consistent and asymptotically normal. Alternatively, one could first estimate θ using standard methods, then plug the estimate of θ in the estimating equations for ψ, and then finally solve those estimating equations for ψ to obtain an estimator for ψ. This leads to the same estimator for ψ as stacking the estimating equations for θ and ψ, because the estimating equations for θ do not involve ψ. Plugging the estimate for θ in the estimating equations for ψ is often easier than solving the stacked estimating equations for θ and ψ in one step.

6. Outline of the proof. Throughout the proof, this article uses fixed versions of $F_{Y^{(t+h)}|\overline{Z}_t}$ satisfying all regularity conditions of Section 7.2. Section 7.3 shows existence of D. It also derives a different expression for D, which is often used in the rest of the proof. Section 7.4 shows existence and uniqueness of solutions $X(t)$ to the differential equation with D, equation (9), with final condition $X(\tau) = Y$.

The proof that this $X(t)$ mimics $Y^{(t)}$ is based on discretization. Section 7.5 therefore considers the situation where the treatment and covariate process Z can be fully described by its values at finitely many fixed times $0 < \tau_1 < \tau_2 < \cdots < \tau_K$ and τ. In fact, this is the discrete-time situation studied in [14], but instead of using the shift-function γ described there as a model this article uses the infinitesimal shift-function D. Proposition 7.8 in Section 7.5 states that in this discrete-time setting with D instead of γ, $X(t)$ mimics $Y^{(t)}$ in the sense that it has the same distribution as $Y^{(t)}$ given the discrete-time $\overline{Z}_t$, under a regularity condition and consistency Assumption 4.1. The proof of Proposition 7.8 is relatively easy, because in this discrete-time setting the continuous solution to the differential equation can be written down explicitly, in terms of conditional distribution functions.

Sections 7.6–7.12 consider the situation where the probability that Z jumps *at* t equals zero for all t. We prove that also in this case, $X(t)$ mimics $Y^{(t)}$, under the conditions of Section 7.2. First, Section 7.6 prepares the discretization by constructing a series $\overline{Z}^{(n)}$, containing more and more information on the covariate and treatment history $\overline{Z}$ as n increases. $\overline{Z}^{(n)}$ depends deterministically on $\overline{Z}$, so that no extra randomness is necessary to construct $\overline{Z}^{(n)}$. The discretization does not change $Y^{(t)}$; just the information on the treatment and covariate process considered is reduced. $\overline{Z}^{(n)}$ is a covariate and treatment history as considered in Section 7.5. Therefore, $D^{(n)}$ can be defined as

$$D^{(n)}\bigl(y, t, \overline{Z}_t^{(n)}\bigr) = \frac{\partial}{\partial h}\bigg|_{h=0} \bigl(F^{-1}_{Y^{(t+h)}|\overline{Z}_t^{(n)}} \circ F_{Y^{(t)}|\overline{Z}_t^{(n)}}\bigr)(y) \tag{11}$$

and we define $X^{(n)}$ as the continuous solution to the differential equation:

$$\frac{d}{dt}X^{(n)}(t) = D^{(n)}\big(X^{(n)}(t), t; \overline{Z}_t^{(n)}\big) \tag{12}$$

with final condition $X^{(n)}(\tau) = Y$. Section 7.7 shows existence of $D^{(n)}$ and provides two expressions for $D^{(n)}$. Section 7.8 shows that the conditions of the discrete-time result are satisfied for the discretized situation, so that Proposition 7.8 guarantees that there exists a continuous solution $X^{(n)}(t)$ to the differential equation (12), with final condition $X^{(n)}(\tau) = Y$ and with the same distribution as $Y^{(t)}$ given $\overline{Z}_t^{(n)}$.

Sections 7.9–7.11 then prove that $X^{(n)}(t)$ converges almost surely to $X(t)$ as n tends to infinity, using a result from differential equation theory which bounds the difference between solutions to differential equations. The proof is concluded in Section 7.12, which shows that $X(t)$ has the same distribution as $Y^{(t)}$ given $\overline{Z}_t$ because $X^{(n)}(t)$ has the same distribution as $Y^{(t)}$ given $\overline{Z}_t^{(n)}$ and $X^{(n)}(t)$ converges almost surely to $X(t)$.

Section 7.13 indicates how the proof can be adapted to include situations where the probability that Z jumps at time t is zero except for at finitely many times t.

7. Proof of main result.

7.1. *Introduction.* The purpose of the current article is to prove that $X(t)$ mimics $Y^{(t)}$. This result is proved in this section for nonsurvival outcomes. Section 7.2 states the assumptions and the precise statement of the result, and Sections 7.3–7.13 provide the proof.

7.2. *Mimicking counterfactual nonsurvival outcomes*: *Assumptions.* This section provides precise conditions under which $X(t)$ mimics $Y^{(t)}$. First, consider the definition of D, equation (1). Notice that D involves an uncountable number of distribution functions $F_{Y^{(t+h)}|\overline{Z}_t}$. In many cases, conditioning on $\overline{Z}_t$ means conditioning on a null-event, so that these conditional distributions are not unique. Every single conditional distribution is almost surely unique (see Lok [10] Section D), but because an uncountable number of them is used (t and h are continuous) this is not sufficient for overall almost sure uniqueness. Therefore, the regularity conditions below should be read as: there exists a collection of conditional distribution functions $F_{Y^{(t+h)}|\overline{Z}_t}$ such that all these regularity conditions are satisfied. These versions of $F_{Y^{(t+h)}|\overline{Z}_t}$ are chosen in the definition of D as well as everywhere else in this article. We only consider $h \geq 0$, so the derivative with respect to h at $h = 0$ is always the right-hand derivative.

With the support of a random variable X, this article means those x such that for every open set U_x containing x, $P(X \in U_x) > 0$. Let y_1 and y_2 be the lower

and upper limit of the support of the outcome of interest Y. In this article, these are assumed to be finite, and moreover, the following is assumed.

ASSUMPTION 7.1 (support).

(a) All $F_{Y^{(t+h)}|\overline{Z}_t}$ for all $t \geq 0$ and for $h \geq 0$ have the same bounded support $[y_1, y_2]$.

(b) All $F_{Y^{(t+h)}|\overline{Z}_t}(y)$ for all $t \geq 0$ and for $h \geq 0$ have a continuous nonzero density $f_{Y^{(t+h)}|\overline{Z}_t}(y)$ on $y \in [y_1, y_2]$.

(c) There exists an $\varepsilon > 0$ such that $f_{Y^{(t)}|\overline{Z}_t}(y) \geq \varepsilon$ for all $y \in [y_1, y_2]$, $\omega \in \Omega$ and $t \in [0, \tau]$.

The support condition may be restrictive for certain applications. Nevertheless, most real-life situations can be approximated this way, since y_1 and y_2 are can have arbitrary (finite) values and $\varepsilon > 0$ can be vary small. Although the support condition may well be stronger than necessary, it simplifies the analysis considerably and, for that reason, it is adopted here.

The remaining regularity conditions are smoothness conditions. They allow for nonsmooth-ness where the covariate and treatment process $Z()$ jumps. This is important since if the covariate and treatment process $Z()$ jumps this can lead to a different prognosis for the person and thus to nonsmoothness of the functions concerned.

ASSUMPTION 7.2 (continuous derivatives). For $\omega \in \Omega$ fixed,

(a) $F_{Y^{(t+h)}|\overline{Z}_t}(y)$ is C^1 in (h, y) for $y \in [y_1, y_2]$ and $h \geq 0$.

(b) If Z does not jump in (t_1, t_2) then both $\frac{\partial}{\partial h}|_{h=0} F_{Y^{(t+h)}|\overline{Z}_t}(y)$ and $\frac{\partial}{\partial y} F_{Y^{(t)}|\overline{Z}_t}(y)$ are continuous in (y, t) on $[y_1, y_2] \times [t_1, t_2)$ and can be continuously extended to $[y_1, y_2] \times [t_1, t_2]$.

Structural nested models in continuous time are meant to estimate the effect of a continuous treatment, for which the effect of a small duration is small. Then Assumption 7.3 is a regularity condition.

ASSUMPTION 7.3 (bounded derivatives).

(a) There exists a constant C_1 such that for all $\omega \in \Omega$, $t, h \geq 0$ and $y \in [y_1, y_2]$,

$$\frac{\partial}{\partial y} F_{Y^{(t+h)}|\overline{Z}_t}(y) \leq C_1.$$

(b) There exists a constant C_2 such that for all $\omega \in \Omega$, $t, h \geq 0$ and $y \in [y_1, y_2]$,

$$\left| \frac{\partial}{\partial h} F_{Y^{(t+h)}|\overline{Z}_t}(y) \right| \leq C_2.$$

ASSUMPTION 7.4 (Lipschitz continuity).

(a) There exists a constant L_1 such that for all $\omega \in \Omega$ and t and $y, z \in [y_1, y_2]$,

$$\left| \frac{\partial}{\partial y} F_{Y^{(t)}|\overline{Z}_t}(y) - \frac{\partial}{\partial z} F_{Y^{(t)}|\overline{Z}_t}(z) \right| \leq L_1 |y - z|.$$

(b) There exists a constant L_2 such that for all $\omega \in \Omega$ and t and $y, z \in [y_1, y_2]$,

$$\left| \frac{\partial}{\partial h}\bigg|_{h=0} F_{Y^{(t+h)}|\overline{Z}_t}(y) - \frac{\partial}{\partial h}\bigg|_{h=0} F_{Y^{(t+h)}|\overline{Z}_t}(z) \right| \leq L_2 |y - z|.$$

Under the regularity conditions in this section and consistency Assumption 4.1, the main result of this article holds, Theorem 4.2: $X(t)$ mimics $Y^{(t)}$ given $\overline{Z}_t$. Thus, under these regularity conditions, if D has the form of equation (3), Theorem 4.2 shows that $X(t)$ mimics $Y^{(t)}$ [or, equation (4) holds], and if D has the form of equation (5), Theorem 4.2 shows that $X(t)$ mimics $Y^{(t)}$ [or, equation (6) holds]. Thus, for equation (4) to hold, equation (2) is not needed, and similarly, for equation (6) to hold, equation (6) with $\sim$ replaced by $=$ is not needed; the form of D is sufficient. This article thus relaxes the assumption of (local) rank preservation.

7.2.1. *Simpler regularity conditions.* I state some more restrictive but simpler conditions implying all the conditions in Section 7.2.

ASSUMPTION 7.5 (regularity condition).

- *(support).*

  (a) There exist finite numbers y_1 and y_2 such that all $F_{Y^{(t+h)}|\overline{Z}_t}$ have the same bounded support $[y_1, y_2]$.

  (b) All $F_{Y^{(t+h)}|\overline{Z}_t}(y)$ have a continuous nonzero density $f_{Y^{(t+h)}|\overline{Z}_t}(y)$ on $y \in [y_1, y_2]$.

  (c) There exists an $\varepsilon > 0$ such that $f_{Y^{(t)}|\overline{Z}_t}(y) \geq \varepsilon$ for all $y \in [y_1, y_2]$, $\omega \in \Omega$ and $t \in [0, \tau]$.

- *(smoothness).* For every $\omega \in \Omega$

  (a) $(y, t, h) \to F_{Y^{(t+h)}|\overline{Z}_t}(y)$ is differentiable with respect to t, y and h with continuous derivatives on $[y_1, y_2] \times [t_1, t_2) \times [0, \infty)$ if Z does not jump in (t_1, t_2), with a continuous extension to $[y_1, y_2] \times [t_1, t_2] \times [0, \infty)$.

  (b) The derivatives of $F_{Y^{(t+h)}|\overline{Z}_t}(y)$ with respect to y and h are bounded by constants C_1 and C_2, respectively.

  (c) $\frac{\partial}{\partial y} F_{Y^{(t)}|\overline{Z}_t}(y)$ and $\frac{\partial}{\partial h}|_{h=0} F_{Y^{(t+h)}|\overline{Z}_t}(y)$ have derivatives with respect to y which are bounded by constants L_1 and L_2, respectively.

7.3. *Existence of and a different expression for D.* The lemma below can be used to prove existence of D and to find a useful formula for D (and later two useful formulas for $D^{(n)}$ in Section 7.5).

LEMMA 7.6. *Suppose that F_h is a family of nondecreasing functions. Suppose that there exists a neighbourhood U_{0,y_0} of $(0, y_0)$ so that $F_h(y)$ is differentiable with respect to y and h on $U_{0,y_0} \cap \{h \geq 0\}$. For $h = 0$, the right-hand derivative is meant. Suppose furthermore that these derivatives are continuous in (h, y). If also $F_0'(y_0)$ is nonzero, then there exists a neighbourhood V_{0,y_0} of $(0, y_0)$ such that on the restriction of this neighbourhood to $h \geq 0$, F_h is invertible. Moreover, $(\frac{\partial}{\partial h} F_h^{-1})(F_h(y))$ exists and satisfies*

$$\frac{\partial}{\partial h} F_h(y) + F_h'(y) \cdot \left(\frac{\partial}{\partial h} F_h^{-1} \right)\big(F_h(y)\big) = 0.$$

PROOF. Define an extension of F to

$$\tilde{U}_{0,y_0} = \big\{(y,h) : h \geq 0 \text{ and } (y,h) \in U_{0,y_0}\big\} \cup \big\{(y,h) : h < 0 \text{ and } (y,-h) \in U_{0,y_0}\big\},$$

an open neighbourhood of $(0, y_0)$, in the following way:

$$\tilde{F}_h(y) = \begin{cases} F_h(y) & \text{if } h \geq 0 \\ 2F_0(y) - F_{-h}(y) & \text{if } h < 0. \end{cases}$$

Define $\phi : U_{h_0,y_0} \to \mathbb{R}^2$ as $\phi(h, y) = (h, \tilde{F}_h(y))$. The result follows from the local inverse function theorem and direct calculation, after noticing that $D(\phi \circ \phi^{-1})$ is the identity mapping; see Lok [10] Section E for details. □

Because of Assumptions 7.2(a) and 7.1(c), Lemma 7.6 can be applied to $F_h(y) = F_{Y^{(t+h)}|\overline{Z}_t}(y)$ with $y_0 = y$. Thus D as defined in equation (1) exists and

$$D(y, t; \overline{Z}_t) = -\frac{\frac{\partial}{\partial h}|_{h=0} F_{Y^{(t+h)}|\overline{Z}_t}(y)}{\frac{\partial}{\partial y} F_{Y^{(t)}|\overline{Z}_t}(y)}. \tag{13}$$

7.4. *Existence and uniqueness of $X(t)$.* This section shows that the differential equation $dX(t)/dt = D(X(t), t; \overline{Z}_t)$ with final condition $X(\tau) = Y$ has a unique continuous solution. Fix ω for the rest of Section 7.4. Since D may be discontinuous at the jump times of the covariate and treatment process Z, we consider the intervals between jumps of Z separately. It suffices to prove existence and uniqueness of $X(t)$ with final condition on any interval between jumps of Z, because with probability one Z only jumps finitely many times.

Hence, suppose that Z does not jump in (t_1, t_2) and that t_1 is either a jump time of Z or 0 and that t_2 is either a jump time of Z or τ. From equation (13), I conclude that $D(y, t; \overline{Z}_t)$ is continuous on $[y_1, y_2] \times [t_1, t_2)$ because of Assumptions 7.2(b)

and 7.1(c). The differential equation has a final condition at the upper end of the interval $[t_1, t_2)$. Therefore, we define $\tilde{D}$ on $[y_1, y_2] \times [t_1, t_2]$ as

$$\tilde{D}(y,t) = \begin{cases} D(y,t;\overline{Z}_t) & \text{if } t \in [t_1, t_2) \\ \lim_{t \uparrow t_2} D(y,t;\overline{Z}_t) & \text{if } t = t_2. \end{cases}$$

This limit exists because of Assumption 7.1(c) and the extension-assumption in Assumption 7.2(b). It makes $\tilde{D}$ continuous on $[y_1, y_2] \times [t_1, t_2]$. When calculating the continuous solution to $dX(t)/dt = D(X(t), t; \overline{Z}_t)$ on $[t_1, t_2]$, one means to use $\tilde{D}$ on $[t_1, t_2]$ if D jumps at t_2.

To prove existence and uniqueness of X on $[t_1, t_2]$, we apply Theorem A.1 to the differential equation with $\tilde{D}$. We check the conditions of Theorem A.1 for $\tilde{D}$. Continuity of $\tilde{D}$ was shown in the previous paragraph. $F^{-1}_{Y^{(t+h)}|\overline{Z}_t} \circ F_{Y^{(t)}|\overline{Z}_t}(y_1) = y_1$ for all h because of Assumption 7.1(a) and (b), so that $D(y_1, t; \overline{Z}_t) = 0$. Similarly, $D(y_2, t; \overline{Z}_t) = 0$. To show that equation (26) holds, notice that global Lipschitz continuity of $\tilde{D}$ in y on $[y_1, y_2] \times [t_1, t_2)$ with Lipschitz constant $C = L_2/\varepsilon + L_1 C_2/\varepsilon^2$ follows from equation (13), since the numerator is bounded by C_2 and is Lipschitz with Lipschiz constant L_2 and also the denominator is Lipschitz with Lipschitz constant L_1 and bounded away from 0 by ε (Assumptions 7.4, 7.3(b) and 7.1(c); see Lok [10] Section F). This same constant works on $[y_1, y_2] \times [t_1, t_2]$ by continuity. By Theorem A.1, the differential equation (9) with $\tilde{D}$ has a unique solution, and this solution stays in $[y_1, y_2]$.

7.5. *Mimicking counterfactual outcomes*: *Discrete time.* This section considers the situation where $\overline{Z}$, the available information on the treatment and covariate process, can be fully described by its values at finitely many fixed-time points $0 = \tau_0 < \tau_1 < \tau_2 < \cdots < \tau_K < \tau_{K+1} = \tau$. At these time points, $Z(t)$ may jump with probability greater than zero. We prove that in this situation, $X(t)$ mimics $Y^{(t)}$.

We assume that there exist conditional distribution functions [1, 16] $F_{Y^{(t)}|\overline{Z}_{\tau_k}}$ satisfying the following regularity condition.

ASSUMPTION 7.7 (smoothness). Suppose that for $k = 0, \ldots, K$ and $t \in [\tau_k, \tau_{k+1}]$ there exist conditional distribution functions $F_{Y^{(t)}|\overline{Z}_{\tau_k}}$ such that:

(a) For all $t \in [\tau_k, \tau_{k+1}]$, $F_{Y^{(t)}|\overline{Z}_{\tau_k}}(y)$ is continuous in y.

(b) For all $t \in [\tau_k, \tau_{k+1}]$, the support of $F_{Y^{(t)}|\overline{Z}_{\tau_k}}(y)$ is an interval.

(c) For $x \in [0, 1]$ fixed, $F^{-1}_{Y^{(t)}|\overline{Z}_{\tau_k}}(x)$ is differentiable with respect to t on $[\tau_k, \tau_{k+1}]$.

Throughout Section 7.5, fixed versions of $F_{Y^{(t)}|\overline{Z}_{\tau_k}}(y)$ are used satisfying Assumption 7.7. Since $\overline{Z}_t$ contains the same information as $\overline{Z}_{\tau_k}$ for $t \in [\tau_k, \tau_{k+1})$, we can and will choose the same versions when conditioning on $\overline{Z}_t$.

PROPOSITION 7.8 (mimicking counterfactual outcomes in discrete time). *Suppose that the treatment and covariate process Z can be fully described by its values at finitely many fixed points $0 = \tau_0 < \tau_1 < \tau_2 < \cdots < \tau_K < \tau_{K+1} = \tau$, and suppose also that smoothness Assumption 7.7 is satisfied. Then $D(y, t; \overline{Z}_t)$ as defined in equation (1) exists for all t. Furthermore if also Assumption 4.1 (consistency) is satisfied, then there exists a continuous solution $X(t)$ to $dX(t)/dt = D(X(t), t; \overline{Z}_t)$ with final condition $X(\tau) = Y$ for which $X(t)$ has the same distribution as $Y^{(t)}$ given $\overline{Z}_t$.*

PROOF. For $t \in [\tau_k, \tau_{k+1})$, $D(y, t; \overline{Z}_t) = \frac{\partial}{\partial h}|_{h=0}(F^{-1}_{Y^{(t+h)}|\overline{Z}_{\tau_k}} \circ F_{Y^{(t)}|\overline{Z}_{\tau_k}})(y)$, so existence of $D(y, t; \overline{Z}_t)$ on each interval $[\tau_k, \tau_{k+1})$ follows from Assumption 7.7(c).

Next, define $\tilde{X}$ as follows. $\tilde{X}(\tau) = Y$, and for $t \in [\tau_k, \tau_{k+1})$ $(k = 0, \ldots, K-1)$,

$$\begin{aligned}\tilde{X}(t) = {} & F^{-1}_{Y^{(t)}|\overline{Z}_{\tau_k}} \circ F_{Y^{(\tau_{k+1})}|\overline{Z}_{\tau_k}} \circ \cdots \circ F^{-1}_{Y^{(\tau_{K-1})}|\overline{Z}_{\tau_{K-1}}} \circ F_{Y^{(\tau_K)}|\overline{Z}_{\tau_{K-1}}} \\ & \circ F^{-1}_{Y^{(\tau_K)}|\overline{Z}_{\tau_K}} \circ F_{Y^{(\tau)}|\overline{Z}_{\tau_K}}(Y).\end{aligned}$$

$\tilde{X}(t)$ is well-defined because of Assumption 7.7(a) and (b). First, we show that $\tilde{X} = X$: it is a continuous solution to $\tilde{X}'(t) = D(\tilde{X}(t), t; \overline{Z}_t)$ with $\tilde{X}(\tau) = Y$. Next, we show that $\tilde{X}(t)$ has the same distribution as $Y^{(t)}$ given $\overline{Z}_t$.

Continuity of $\tilde{X}$ on $[\tau_k, \tau_{k+1})$ is clear from Assumption 7.7(c). Moreover,

$$\begin{aligned}\lim_{t \uparrow \tau_{k+1}} \tilde{X}(t) &= \lim_{t \uparrow \tau_{k+1}} F^{-1}_{Y^{(t)}|\overline{Z}_{\tau_k}} \circ F_{Y^{(\tau_{k+1})}|\overline{Z}_{\tau_k}}\big(\tilde{X}(\tau_{k+1})\big) \\ &= F^{-1}_{Y^{(\tau_{k+1})}|\overline{Z}_{\tau_k}} \circ F_{Y^{(\tau_{k+1})}|\overline{Z}_{\tau_k}}\big(\tilde{X}(\tau_{k+1})\big)\end{aligned}$$

because of Assumption 7.7(c), which is equal to $\tilde{X}(\tau_{k+1})$ because of Assumption 7.7(b). Thus, $\tilde{X}(t)$ is also continuous from the left at $t = \tau_{k+1}$. For $t \in [\tau_k, \tau_{k+1})$, $\tilde{X}$ satisfies the differential equation:

$$\begin{aligned}\tilde{X}'(t) &= \frac{\partial}{\partial h}\bigg|_{h=0} F^{-1}_{Y^{(t+h)}|\overline{Z}_{\tau_k}} \circ F_{Y^{(\tau_{k+1})}|\overline{Z}_{\tau_k}}\big(\tilde{X}(\tau_{k+1})\big) \\ &= \left(\frac{\partial}{\partial h}\bigg|_{h=0} F^{-1}_{Y^{(t+h)}|\overline{Z}_t} \circ F_{Y^{(t)}|\overline{Z}_t}\right) \circ F^{-1}_{Y^{(t)}|\overline{Z}_{\tau_k}} \circ F_{Y^{(\tau_{k+1})}|\overline{Z}_{\tau_k}}\big(\tilde{X}(\tau_{k+1})\big) \\ &= D\big(\tilde{X}(t), t; \overline{Z}_t\big),\end{aligned}$$

where in the second line it is used that conditioning on $\overline{Z}_t$ is the same as conditioning on $\overline{Z}_{\tau_k}$, so that $F_{Y^{(t)}|\overline{Z}_t} \circ F^{-1}_{Y^{(t)}|\overline{Z}_{\tau_k}}$ is the identity because of Assumption 7.7(a) and (b). Thus, indeed $\tilde{X}$ is a continuous solution to $\tilde{X}'(t) = D(\tilde{X}(t), t; \overline{Z}_t)$ with $\tilde{X}(\tau) = Y$.

Next, we prove that $\tilde{X}(t)$ has the same distribution as $Y^{(t)}$ given $\overline{Z}_t$ by induction, starting at $t = \tau$, then $t \in [\tau_K, \tau)$, etcetera. For $t = \tau$, $\tilde{X}(\tau) = Y$, so that $\tilde{X}(\tau)$

has the same distribution as $Y^{(\tau)}$ given $\overline{Z}_\tau$ because of Assumption 4.1. For the induction step, suppose that for $t \in [\tau_k, \tau]$ (for $k = K+1$ read $t = \tau$), $\tilde{X}(t)$ has the same distribution as $Y^{(t)}$ given $\overline{Z}_t$. Thus, $\tilde{X}(\tau_k)$ has the same distribution as $Y^{(\tau_k)}$ given $\overline{Z}_{\tau_k}$, and hence $\tilde{X}(\tau_k)$ also has the same distribution as $Y^{(\tau_k)}$ given $\overline{Z}_{\tau_{k-1}}$. Therefore, Assumption 7.7(a) implies that $F_{Y^{(\tau_k)}|\overline{Z}_{\tau_{k-1}}}(\tilde{X}(\tau_k))$ is uniformly distributed on $[0,1]$ given $\overline{Z}_{\tau_{k-1}}$ (Lemma D.8 has a formal proof). Then $\tilde{X}(t) = F^{-1}_{Y^{(t)}|\overline{Z}_{\tau_{k-1}}} \circ F_{Y^{(\tau_k)}|\overline{Z}_{\tau_{k-1}}}(\tilde{X}(\tau_k))$ has distribution function $F_{Y^{(t)}|\overline{Z}_{\tau_{k-1}}}$ given $\overline{Z}_{\tau_{k-1}}$ (Lemma D.9 has a formal proof), so also given $\overline{Z}_t$. That finishes the induction step, so that indeed $\tilde{X}(t)$ mimics $Y^{(t)}$ for all $t \in [0, \tau]$. □

7.6. *Discretization and choices of conditional distributions*. We return to the continuous-time setting and define a discretization of the covariate and treatment process Z. Later, we will apply the result of the previous section to this discretized continuous-time setting. This section also chooses versions of the conditional distribution functions given this discretized process.

For n fixed define $\tau_0^{(n)} = 0$, $\tau_1^{(n)} = \frac{1}{2^n}\tau$, $\tau_2^{(n)} = \frac{2}{2^n}\tau$, …, $\tau_{2^n}^{(n)} = \frac{2^n}{2^n}\tau = \tau$. Consider the grid at stage n consisting of these points. This way the interval $[0, \tau]$ is split up into 2^n intervals of equal length, and when n increases points are added in the middle of these intervals. For ease of notation, the superscript $^{(n)}$ in $\tau_k^{(n)}$ is dropped if it is clear which n is meant. Define $\overline{Z}_t^{(n)} = (Z(\tau_k^{(n)}) : 0 \le \tau_k^{(n)} \le t)$ if Z takes values in a discrete space, $\overline{Z}_t^{(n)} = (1_{[\frac{i}{2^n}, \frac{i+1}{2^n})}(Z(\tau_k^{(n)})) : 0 \le \tau_k^{(n)} \le t, i \in \mathbb{Z})$ if Z takes values in $\mathbb{R}$ and $\overline{Z}_t^{(n)} = (1_{[\frac{i}{2^n}, \frac{i+1}{2^n})}(Z(\tau_k^{(n)})_j) : 0 \le \tau_k^{(n)} \le t, i \in \mathbb{Z}, j = 1, \dots, m)$ if Z takes values in $\mathbb{R}^m$.

With this discretization, the information about $\overline{Z}_t$ contained in $\overline{Z}_t^{(n)}$ increases with n: once a grid point is added it stays on the grid for n larger, and the information about Z in a fixed grid point also increases with n. Note also that $\overline{Z}_t^{(n)}$ depends deterministically on $\overline{Z}_t$, so that no extra randomness is necessary to construct $\overline{Z}_t^{(n)}$. Thus, $\overline{Z}_t^{(n)}$ has the properties promised in the outline of the proof, Section 6.

Next, versions of conditional distributions are chosen. Recall $\overline{Z}_{\tau_k}$ takes values in the space of cadlag functions on $[0, \tau_k]$ with the projection σ-algebra, which is the same as the Skorohod-σ-algebra ([2] Theorem 14.5). This space is Polish ([2] Chapter 3). Therefore, there exists a conditional distribution $P_{\overline{Z}_{\tau_k}|\overline{Z}^{(n)}_{\tau_k}}$ ([1] Section 10.3 or [16]). Moreover,

$$P\big(Y^{(t+h)} \le y | \overline{Z}^{(n)}_{\tau_k}\big) = \int F_{Y^{(t+h)}|\overline{Z}_{\tau_k}=z}(y)\, dP_{\overline{Z}_{\tau_k}|\overline{Z}^{(n)}_{\tau_k}}(z) \qquad \text{a.s.}$$

This is a conditional distribution function: it is nondecreasing in y since all $F_{Y^{(t+h)}|\overline{Z}_{\tau_k}=z}(y)$ are nondecreasing because they are conditional distribution functions, and because of Lebesgue's dominated convergence theorem the limit for

$y \to -\infty$ equals 0 and the limit for $y \to \infty$ equals 1. Therefore, the following choices can be made.

NOTATION 7.9. We choose fixed conditional distributions $P_{\overline{Z}_{\tau_k}|\overline{Z}^{(n)}_{\tau_k}}$. I also choose

$$F_{Y^{(t)}|\overline{Z}^{(n)}_{\tau_k}}(y) = \int F_{Y^{(t)}|\overline{Z}_{\tau_k}=z}(y)\, dP_{\overline{Z}_{\tau_k}|\overline{Z}^{(n)}_{\tau_k}}(z),$$

with $F_{Y^{(t)}|\overline{Z}_{\tau_k}=z}$ as in Section 7.2, to be the version of the conditional distribution function of $Y^{(t)}$ given $\overline{Z}^{(n)}_{\tau_k}$ which is used in the rest of the proof. If $s \in (\tau_k, \tau_{k+1})$, the same version for $F_{Y^{(t)}|\overline{Z}^{(n)}_s}$ is chosen; this is possible since for $s \in (\tau_k, \tau_{k+1})$, $\overline{Z}^{(n)}_s = \overline{Z}^{(n)}_{\tau_k}$.

Notice that $Z^{(n)}$ has been constructed with values in a discrete space. This will assure that the two different expressions for $D^{(n)}$ in Section 7.7 below are equal except for at a null set which does not depend on y and t.

7.7. *Existence of and two expressions for $D^{(n)}$.* This section proves existence of $D^{(n)}$ as defined in equation (11), Section 6. Moreover, two useful formulas for $D^{(n)}$ are proven. One is used to prove smoothness of $D^{(n)}$, the other formula is used to prove that $D^{(n)}$ converges to D.

First, existence of $D^{(n)}$ is shown. Fix n and t, and choose $\tau^{(n)}_k$ such that $t \in [\tau^{(n)}_k, \tau^{(n)}_{k+1})$. Define

$$F_h(y) = F_{Y^{(t+h)}|\overline{Z}^{(n)}_{\tau_k}}(y) = \int F_{Y^{(t+h)}|\overline{Z}_{\tau_k}=z}(y)\, dP_{\overline{Z}_{\tau_k}|\overline{Z}^{(n)}_{\tau_k}}(z).$$

To apply Lemma 7.6 on $F_h(y)$, in $(h_0, y_0) = (0, y)$, we check the conditions. Clearly, $F_h(y)$ is nondecreasing. We show that $F_h(y)$ is differentiable with respect to y with derivative $\int \frac{\partial}{\partial y} F_{Y^{(t+h)}|\overline{Z}_{\tau_k}=z}(y)\, dP_{\overline{Z}_{\tau_k}|\overline{Z}^{(n)}_{\tau_k}}(z)$. For ω fixed, $P_{\overline{Z}_{\tau_k}|\overline{Z}^{(n)}_{\tau_k}}$ is a probability measure on $\overline{\mathcal{Z}}_{\tau_k}$. Moreover, $\frac{\partial}{\partial y} F_{Y^{(t+h)}|\overline{Z}_{\tau_k}=z}(y)$ is bounded by C_1, which is integrable with respect to $P_{\overline{Z}_{\tau_k}|\overline{Z}^{(n)}_{\tau_k}}$, and also $F_{Y^{(t+h)}|\overline{Z}_{\tau_k}=z}(y)$ is integrable with respect to $P_{\overline{Z}_{\tau_k}|\overline{Z}^{(n)}_{\tau_k}}$, since bounded by 1. Therefore, $F_h(y)$ is differentiable with respect to y with derivative $\int \frac{\partial}{\partial y} F_{Y^{(t+h)}|\overline{Z}_{\tau_k}=z}(y)\, dP_{\overline{Z}_{\tau_k}|\overline{Z}^{(n)}_{\tau_k}}(z)$. With the same reasoning [but with Assumption 7.3(b) instead of 7.3(a)], $F_h(y)$ is differentiable with respect to h with derivative $\int \frac{\partial}{\partial h} F_{Y^{(t+h)}|\overline{Z}_{\tau_k}=z}(y)\, dP_{\overline{Z}_{\tau_k}|\overline{Z}^{(n)}_{\tau_k}}(z)$. That these derivatives of $F_h(y)$ with respect to y and h are continuous in (y, h) follows from Lebesgue's dominated convergence theorem applied on the expressions we just derived [the conditions are satisfied because of Assumptions 7.2(a) and

7.3]. Furthermore, $F_0'(y) = \int \frac{\partial}{\partial y} F_{Y^{(t)}|\overline{Z}_{\tau_k}=z}(y)\, dP_{\overline{Z}_{\tau_k}|\overline{Z}^{(n)}_{\tau_k}}(z)$ is nonzero [Assumption 7.1(b)]. Thus, the conditions of Lemma 7.6 are satisfied for $F_h(y)$, and therefore $\frac{\partial}{\partial h} F^{-1}_{Y^{(t+h)}|\overline{Z}^{(n)}_{\tau_k}}(y)$ exists, and $D^{(n)}(y,t;\overline{Z}^{(n)}_t)$ exists and satisfies

$$
\begin{aligned}
D^{(n)}\big(y,t;\overline{Z}^{(n)}_t\big) &= \frac{\partial}{\partial h}\bigg|_{h=0} \big(F^{-1}_{Y^{(t+h)}|\overline{Z}^{(n)}_{\tau_k}} \circ F_{Y^{(t)}|\overline{Z}^{(n)}_{\tau_k}}\big)(y) \\
&= -\frac{\frac{\partial}{\partial h}|_{h=0} \int F_{Y^{(t+h)}|\overline{Z}_{\tau_k}=z}(y)\, dP_{\overline{Z}_{\tau_k}|\overline{Z}^{(n)}_{\tau_k}}(z)}{\frac{\partial}{\partial y} \int F_{Y^{(t)}|\overline{Z}_{\tau_k}=z}(y)\, dP_{\overline{Z}_{\tau_k}|\overline{Z}^{(n)}_{\tau_k}}(z)} \qquad (14) \\
&= -\frac{\int \frac{\partial}{\partial h}|_{h=0} F_{Y^{(t+h)}|\overline{Z}_{\tau_k}=z}(y)\, dP_{\overline{Z}_{\tau_k}|\overline{Z}^{(n)}_{\tau_k}}(z)}{\int \frac{\partial}{\partial y} F_{Y^{(t)}|\overline{Z}_{\tau_k}=z}(y)\, dP_{\overline{Z}_{\tau_k}|\overline{Z}^{(n)}_{\tau_k}}(z)}.
\end{aligned}
$$

Next, the second expression for $D^{(n)}$ is derived. We show that there exists an $\Omega' \subset \Omega$ with probability one such that

$$
D^{(n)}\big(y,t;\overline{Z}^{(n)}_t\big) = -\frac{E[\frac{\partial}{\partial h}|_{h=0} F_{Y^{(t+h)}|\overline{Z}_t}(y)|\overline{Z}^{(n)}_t]}{E[\frac{\partial}{\partial y} F_{Y^{(t)}|\overline{Z}_t}(y)|\overline{Z}^{(n)}_t]} \qquad \forall \omega \in \Omega' \forall y \forall t \forall n. \tag{15}
$$

First, we choose this Ω', in such a way that on Ω' conditional probabilities given $\overline{Z}^{(n)}_{\tau_k}$ are unique, for all n and τ_k. Fix n and τ_k for a moment. It is known from general theory about conditioning that conditional probabilities given $\overline{Z}^{(n)}_{\tau_k} = z$ can be written as a measurable function of z. It is also known that conditional probabilities given $\overline{Z}^{(n)}_{\tau_k} = z$ are almost surely unique. Combining these two facts, it follows that conditional probabilities given $\overline{Z}^{(n)}_{\tau_k} = z$ are unique except for at ω's for which $\overline{Z}^{(n)}_{\tau_k}(\omega)$ has probability zero, that is, except for ω's in

$$
\bigcup_{z: P(\overline{Z}^{(n)}_{\tau_k}=z)=0} \big\{\omega \in \Omega : \overline{Z}^{(n)}_{\tau_k}(\omega) = z\big\}.
$$

Since, by construction, $\overline{Z}^{(n)}_{\tau_k}$ takes only countably many values, this is a countable union of null sets and thus a null set. Define

$$
\Omega' = \Omega \setminus \bigcup_{n\in\mathbb{N}} \bigcup_{k\in\{0,\ldots,2^n\}} \bigcup_{z: P(\overline{Z}^{(n)}_{\tau_k}=z)=0} \big\{\omega \in \Omega : \overline{Z}^{(n)}_{\tau_k}(\omega) = z\big\}. \tag{16}
$$

This set has probability one since its complement is a countable union of null sets: $\mathbb{N}$ is countable and for each n there are only finitely many k. On this Ω' conditional probabilities given $\overline{Z}^{(n)}_{\tau_k}$ are unique, for all n and τ_k.

Next, it is shown that equation (15) holds for Ω' as defined in equation (16). As shown in Section 7.6, there exists a conditional distribution $P_{\overline{Z}_t|\overline{Z}^{(n)}_{\tau_k}}$. For $t \geq \tau_k$ and $h \geq 0$, $F_h(y) := P(Y^{(t+h)} \leq y|\overline{Z}^{(n)}_{\tau_k}) = \int F_{Y^{(t+h)}|\overline{Z}_t=z}(y)\, dP_{\overline{Z}_t|\overline{Z}^{(n)}_{\tau_k}}(z)$ a.s. On Ω' this version is the same as the one used in the definition of $D^{(n)}$ of equation (11), since conditional probabilities given $\overline{Z}^{(n)}_{\tau_k}$ are unique on Ω'. Verifying the conditions of Lemma 7.6 can be done in exactly the same way as for the first expression for $D^{(n)}$. Therefore, Lemma 7.6 implies that for $\omega \in \Omega'$ and $t \in [\tau_k, \tau_{k+1})$,

$$\begin{aligned} D^{(n)}(y, t; \overline{Z}^{(n)}_t) &= \frac{\partial}{\partial h}\Big|_{h=0} (F^{-1}_{Y^{(t+h)}|\overline{Z}^{(n)}_{\tau_k}} \circ F_{Y^{(t)}|\overline{Z}^{(n)}_{\tau_k}})(y) \\ &= -\frac{\frac{\partial}{\partial h}|_{h=0} \int F_{Y^{(t+h)}|\overline{Z}_t=z}(y)\, dP_{\overline{Z}_t|\overline{Z}^{(n)}_{\tau_k}}(z)}{\frac{\partial}{\partial y} \int F_{Y^{(t)}|\overline{Z}_t=z}(y)\, dP_{\overline{Z}_t|\overline{Z}^{(n)}_{\tau_k}}(z)} \\ &= -\frac{E[\frac{\partial}{\partial h}|_{h=0} F_{Y^{(t+h)}|\overline{Z}_t}(y)|\overline{Z}^{(n)}_{\tau_k}]}{E[\frac{\partial}{\partial y} F_{Y^{(t)}|\overline{Z}_t}(y)|\overline{Z}^{(n)}_{\tau_k}]}. \end{aligned}$$

Equation (15) follows.

7.8. *Applying the discrete-time result.*

LEMMA 7.10. *Suppose that regularity conditions 7.1–7.4 and consistency Assumption 4.1 are satisfied. Then for every n there exists a continuous solution $X^{(n)}(t)$ to the differential equation with $D^{(n)}$ with final condition $X^{(n)}(\tau) = Y$. $X^{(n)}(t)$ is unique on Ω' of equation (16). Furthermore, $X^{(n)}(t)$ has the same conditional distribution as $Y^{(t)}$ given $\overline{Z}^{(n)}_t$.*

PROOF. Fix n. First, we show that there exists a continuous solution $X^{(n)}$ for which $X^{(n)}(t)$ has the same conditional distribution as $Y^{(t)}$ given $\overline{Z}^{(n)}_t$, using Proposition 7.8. Thus, we check that the conditional distributions $F_{Y^{(t)}|\overline{Z}^{(n)}_{\tau_k}}$ of $Y^{(t)}$ given $\overline{Z}^{(n)}_{\tau_k}$ chosen in Notation 7.9 satisfy Assumption 7.7. In the second paragraph of Section 7.7, we showed that $F_{Y^{(t)}|\overline{Z}^{(n)}_{\tau_k}}(y)$ is strictly increasing and differentiable with respect to y on $[y_1, y_2]$, which accounts for Assumption 7.7(a) and (b). Just before equation (14), it was concluded that for $x \in [0, 1]$ fixed, $F^{-1}_{Y^{(t)}|\overline{Z}^{(n)}_{\tau_k}}(x)$ is differentiable with respect to t on $[\tau_k, \tau_{k+1}]$, which accounts for Assumption 7.7(c). Hence, Proposition 7.8 guarantees existence of a continuous solution $X^{(n)}$ to $X^{(n)}(t)' = D^{(n)}(X^{(n)}(t), t)$ with final condition $X^{(n)}(\tau) = Y$ and with $X^{(n)}(t) \sim Y^{(t)}$ given $\overline{Z}^{(n)}_t$.

Proposition 7.8 does not imply that $X^{(n)}$ is unique. Almost sure uniqueness of $X^{(n)}$ follows with Theorem A.1 in the Appendix along the same lines as uniqueness of X (see Section 7.4), but using equations (14) and (15) for $D^{(n)}$ instead of equation (13) for D, as follows. Fix n and suppose that $t \in [\tau_k, \tau_{k+1})$. First, it is proven that $D^{(n)}$ is continuous on $[y_1, y_2] \times [\tau_k, \tau_{k+1})$ with a continuous extension to $[y_1, y_2] \times [\tau_k, \tau_{k+1}]$, using equation (14). Expression (14) for $D^{(n)}$ has an obvious extension $\tilde{D}^{(n)}$ to $[\tau_k, \tau_{k+1}]$. We prove that this $\tilde{D}^{(n)}$ is continuous on $[y_1, y_2] \times [\tau_k, \tau_{k+1}]$. To show that $\int \frac{\partial}{\partial h}|_{h=0} F_{Y^{(t+h)}|\overline{Z}_{\tau_k}=z}(y)\, dP_{\overline{Z}_{\tau_k}|\overline{Z}^{(n)}_{\tau_k}}(z)$ and $\int \frac{\partial}{\partial y} F_{Y^{(t)}|\overline{Z}_{\tau_k}=z}(y)\, dP_{\overline{Z}_{\tau_k}|\overline{Z}^{(n)}_{\tau_k}}(z)$ are continuous in (y, t) Lebesgue's dominated convergence theorem can be used, as follows:

$$\frac{\partial}{\partial h}\bigg|_{h=0} F_{Y^{(t+h)}|\overline{Z}_{\tau_k}=z}(y) = \frac{\partial}{\partial h}\bigg|_{h=t-\tau_k} F_{Y^{(\tau_k+h)}|\overline{Z}_{\tau_k}=z}(y)$$

and

$$\frac{\partial}{\partial y} F_{Y^{(t)}|\overline{Z}_{\tau_k}=z}(y) = \frac{\partial}{\partial y} F_{Y^{(\tau_k+(t-\tau_k))}|\overline{Z}_{\tau_k}=z}(y)$$

are continuous in (y, t) because of Assumption 7.2(a). Both these derivatives are bounded because of Assumption 7.3. Therefore, Lebesgue's dominated convergence theorem implies that the integrals of these derivatives with respect to the measure $\mu = P_{\overline{Z}_{\tau_k}|\overline{Z}^{(n)}_{\tau_k}}$ are continuous in (y, t). Because of Assumption 7.1(b) the denominator of $\tilde{D}^{(n)}$ is nonzero for $y \in [y_1, y_2]$, so that indeed $\tilde{D}^{(n)}$ is continuous in (y, t) on $[y_1, y_2] \times [\tau_k, \tau_{k+1}]$.

Next, it is shown that $D^{(n)}$ is Lipschitz continuous in y on $[y_1, y_2] \times [\tau_k, \tau_{k+1}]$ with Lipschitz constant $L_2/\varepsilon + C_2 L_1/\varepsilon^2$ for all $\omega \in \Omega'$, with Ω' as in equation (16). Expression (14) for $D^{(n)}$ on Ω' has an obvious extension $\tilde{D}^{(n)}$ to $[\tau_k, \tau_{k+1}]$. That this $\tilde{D}^{(n)}$ is Lipschitz continuous in y on $[y_1, y_2] \times [\tau_k, \tau_{k+1}]$ with Lipschitz constant $L_2/\varepsilon + C_2 L_1/\varepsilon^2$ on Ω' follows the same way as for D in Section 7.4. Because of Assumption 7.1(c), the denominator is bounded away from 0 for $y \in [y_1, y_2]$, and because of Assumption 7.1(a), the numerator is equal to zero for $y = y_1$ and for $y = y_2$. Hence, on Ω', $\tilde{D}^{(n)}(y_1, t) = \tilde{D}^{(n)}(y_2, t) = 0$. Therefore, Theorem A.1 implies that, on Ω', there exists a unique solution to the differential equation with $\tilde{D}^{(n)}$ on $[\tau_k, \tau_{k+1}]$, and this solution stays in $[y_1, y_2]$. Since for n fixed there are only finitely many τ_k, the same is true on $[0, \tau]$. □

7.9. *Bounding the difference between X and $X^{(n)}$ in terms of D and $D^{(n)}$.* To bound the difference between X and $X^{(n)}$ in terms of D and $D^{(n)}$, Theorem A.1 is applied on $y = X(t)$ and $z = X^{(n)}(t)$. Since we need that both D and $D^{(n)}$ are continuous, we apply Theorem A.1 on the intervals between the jumps of Z and the grid points $\tau_k^{(n)}$. Fix n and restrict ω to $\omega \in \Omega'$, with Ω' the set of probability one as defined in equation (16), so that the expression for $D^{(n)}$ of equation (15)

can be used. The bound will thus hold almost surely. To focus attention on the differential equations, the $\overline{Z}_t$'s and $\overline{Z}_t^{(n)}$'s in D and $D^{(n)}$ are skipped below.

Suppose that (t_1, t_2) is such an interval including no jumps of Z and no grid points at stage n. We check the conditions of Theorem A.1 for $y = X(t)$ and $z = X^{(n)}(t)$. Section 7.4 already showed that $D : [y_1, y_2] \times [t_1, t_2) \to \mathbb{R}$ has a continuous extension $\tilde{D} : [y_1, y_2] \times [t_1, t_2] \to \mathbb{R}$ which satisfies the conditions of Theorem A.1, with C the constant function $L_2/\varepsilon + C_2 L_1/\varepsilon^2$, and in the proof of Lemma 7.10 in Section 7.8, it was shown that on Ω' the same is true for $D^{(n)}$. Therefore, Theorem A.1 implies that for $t \in [t_1, t_2]$, with $C = L_2/\varepsilon + C_2 L_1/\varepsilon^2$ as above,

$$
\begin{aligned}
\bigl|X^{(n)}(t) - X(t)\bigr| &\le e^{\int_t^{t_2} C\,ds}\bigl|X^{(n)}(t_2) - X(t_2)\bigr| \\
&\quad + \int_t^{t_2} e^{\int_t^s C\,d\eta}\bigl|D\bigl(X^{(n)}(s), s\bigr) - D^{(n)}\bigl(X^{(n)}(s), s\bigr)\bigr|\,ds \\
&= e^{C\cdot(t_2 - t)}\bigl|X^{(n)}(t_2) - X(t_2)\bigr| \\
&\quad + \int_t^{t_2} e^{C\cdot(s-t)}\bigl|D\bigl(X^{(n)}(s), s\bigr) - D^{(n)}\bigl(X^{(n)}(s), s\bigr)\bigr|\,ds.
\end{aligned} \tag{17}
$$

If Z does not jump in $[(1 - 1/2^n)\tau, \tau]$, (17) can be applied on $[(1 - 1/2^n)\tau, \tau]$, and since $X^{(n)}(\tau) = X(\tau) = Y$ it follows that on $[(1 - 1/2^n)\tau, \tau]$,

$$
\bigl|X^{(n)}(t) - X(t)\bigr| \le \int_t^{\tau} e^{C\cdot(s-t)}\bigl|D\bigl(X^{(n)}(s), s\bigr) - D^{(n)}\bigl(X^{(n)}(s), s\bigr)\bigr|\,ds. \tag{18}
$$

If Z does not jump after $(1 - 2/2^n)\tau$, one can also apply (17) on $[(1 - 2/2^n)\tau, (1 - 1/2^n)\tau]$, and using equation (18) for $t = (1 - 1/2^n)\tau$, it follows that equation (18) also holds on $[(1 - 2/2^n)\tau, (1 - 1/2^n)\tau]$:

$$
\begin{aligned}
&\bigl|X^{(n)}(t) - X(t)\bigr| \\
&\quad \le e^{C\cdot((1-\frac{1}{2^n})\tau - t)} \int_{(1-\frac{1}{2^n})\tau}^{\tau} e^{C\cdot(s-(1-\frac{1}{2^n})\tau)}\bigl|D\bigl(X^{(n)}(s), s\bigr) \\
&\qquad - D^{(n)}\bigl(X^{(n)}(s), s\bigr)\bigr|\,ds \\
&\qquad + \int_t^{(1-\frac{1}{2^n})\tau} e^{C\cdot(s-t)}\bigl|D\bigl(X^{(n)}(s), s\bigr) - D^{(n)}\bigl(X^{(n)}(s), s\bigr)\bigr|\,ds \\
&\quad = \int_t^{\tau} e^{C\cdot(s-t)}\bigl|D\bigl(X^{(n)}(s), s\bigr) - D^{(n)}\bigl(X^{(n)}(s), s\bigr)\bigr|\,ds.
\end{aligned}
$$

If Z does not jump in $((1 - m/2^n)\tau, \tau]$ and $t \in ((1 - m/2^n)\tau, \tau]$ then, with the same reasoning, equation (18) holds on $t \in ((1 - m/2^n)\tau, \tau]$. Suppose now that Z jumps in $((1 - (m+1)/2^n)\tau, (1 - m/2^n)\tau]$. Then this interval can be split up into the part before and the part after the jump, so that, again with the same reasoning as before and since both $X^{(n)}$ and X are continuous in t, equation (18) still holds.

With probability one, there are at most finitely many jump times of Z, so that equation (18) holds almost surely for all t, and even

$$\begin{aligned}
&\sup_{t\in[0,\tau]} \bigl|X^{(n)}(t) - X(t)\bigr| \\
&\quad\le \sup_{t\in[0,\tau]} \int_t^\tau e^{C\cdot(s-t)} \bigl|D\bigl(X^{(n)}(s), s\bigr) - D^{(n)}\bigl(X^{(n)}(s), s\bigr)\bigr|\, ds \qquad (19)\\
&\quad= \int_0^\tau e^{C\cdot s} \bigl|D\bigl(X^{(n)}(s), s\bigr) - D^{(n)}\bigl(X^{(n)}(s), s\bigr)\bigr|\, ds \qquad \text{a.s.}
\end{aligned}$$

7.10. *Convergence of $D^{(n)}$ to D.* This section proves that $D^{(n)}(y,t;\overline{Z}_t^{(n)})$ converges almost surely to $D(y,t;\overline{Z}_t)$, for fixed $(y,t)\in[y_1,y_2]\times[0,\tau]$. From equations (13) and (15), it follows that

$$D(y,t;\overline{Z}_t) = -\frac{\frac{\partial}{\partial h}|_{h=0} F_{Y^{(t+h)}|\overline{Z}_t}(y)}{\frac{\partial}{\partial y} F_{Y^{(t)}|\overline{Z}_t}(y)}$$

and

$$D^{(n)}\bigl(y,t;\overline{Z}_t^{(n)}\bigr) = -\frac{E[\frac{\partial}{\partial h}|_{h=0} F_{Y^{(t+h)}|\overline{Z}_t}(y)|\overline{Z}_t^{(n)}]}{E[\frac{\partial}{\partial y} F_{Y^{(t)}|\overline{Z}_t}(y)|\overline{Z}_t^{(n)}]} \qquad \text{a.s.}$$

Lévy's upward theorem (see, e.g., [29] p. 134) can be applied to the denominator and the numerator of $D^{(n)}$, since both $\frac{\partial}{\partial h}|_{h=0} F_{Y^{(t+h)}|\overline{Z}_t}(y)$ and $\frac{\partial}{\partial y} F_{Y^{(t)}|\overline{Z}_t}(y)$ are bounded (Assumption 7.3). Lévy's upward theorem leads to

$$E\left[\frac{\partial}{\partial h}\bigg|_{h=0} F_{Y^{(t+h)}|\overline{Z}_t}(y)\Big|\overline{Z}_t^{(n)}\right] \to E\left[\frac{\partial}{\partial h}\bigg|_{h=0} F_{Y^{(t+h)}|\overline{Z}_t}(y)\Big|\sigma\left(\bigcup_{n=1}^{\infty}\overline{Z}_t^{(n)}\right)\right] \qquad \text{a.s.}$$

and

$$E\left[\frac{\partial}{\partial y} F_{Y^{(t)}|\overline{Z}_t}(y)\Big|\overline{Z}_t^{(n)}\right] \to E\left[\frac{\partial}{\partial y} F_{Y^{(t)}|\overline{Z}_t}(y)\Big|\sigma\left(\bigcup_{n=1}^{\infty}\overline{Z}_t^{(n)}\right)\right] \qquad \text{a.s.}$$

as $n\to\infty$. The conditioning on $\sigma(\bigcup_{n=1}^{\infty}\overline{Z}_t^{(n)})$ can be replaced by conditioning on $\overline{Z}_t$ in both expressions, because of Lemma A.2 in the Appendix. Since moreover the denominators are bounded away from 0 [Assumption 7.1(c)], the continuous mapping theorem implies that, for fixed $(y,t)\in[y_1,y_2]\times[0,\tau]$,

$$D^{(n)}\bigl(y,t;\overline{Z}_t^{(n)}\bigr) \to D(y,t;\overline{Z}_t) \qquad \text{a.s.} \qquad (20)$$

7.11. *$X^{(n)}(t)$ converges to $X(t)$ and $X(t)$ is measurable.* To show that $X^{(n)}(t)$ converges almost surely to $X(t)$ and that $X(t)$ is measurable, the bound of equation (19) and almost sure convergence of $D^{(n)}(y,t)$ to $D(y,t)$ for (y,t) fixed of equation (20) are the starting point.

First, it is proven that for s fixed, $D^{(n)}(X^{(n)}(s),s) - D(X^{(n)}(s),s)$ converges almost surely to 0. Recall from Section 7.4 that $D : [y_1,y_2] \times [t_1,t_2) \to \mathbb{R}$ has a continuous extension $\tilde{D} : [y_1,y_2] \times [t_1,t_2] \to \mathbb{R}$ which is Lipschitz continuous in y with Lipschitz constant $L_2/\varepsilon + C_2L_1/\varepsilon^2$. Recall also that in the proof of Lemma 7.10 in Section 7.8 it was shown that on Ω', the set of probability one of equation (16), the same is true for $D^{(n)}$. Therefore, the pointwise almost sure convergence of $D^{(n)}(y,t)$ to $D(y,t)$ of equation (20) implies that for fixed s indeed

$$\left|D^{(n)}\big(X^{(n)}(s),s\big) - D\big(X^{(n)}(s),s\big)\right| \to 0 \qquad \text{a.s.} \tag{21}$$

(for details, see Lok [10] Section H).

To show that equation (21) implies that the bound of (19) converges almost surely to 0, define

$$A = \left\{(s,\omega) \in [0,\tau] \times \Omega : \left|D^{(n)}\big(X^{(n)}(s),s\big) - D\big(X^{(n)}(s),s\big)\right| \to 0\right\},$$

with A_s its section at s and A_ω its section at ω. Then

$$A_s = \left\{\omega \in \Omega : \left|D^{(n)}\big(X^{(n)}(s),s\big) - D\big(X^{(n)}(s),s\big)\right| \to 0\right\}$$

has probability one because equation (21). Therefore, using Fubini's theorem, with λ the Lebesgue-measure on $[0,\tau]$,

$$\begin{aligned}(\lambda \times P)(A) &= \int_{(0,\tau)} P(A_s)\, d\lambda(s) \\ &= \int_{(0,\tau)} 1\, d\lambda(s) = \tau.\end{aligned}$$

Also, by Fubini's theorem,

$$(\lambda \times P)(A) = \int \lambda(A_\omega)\, dP(\omega),$$

so that since $\lambda(A_\omega) \le \tau$, $\lambda(A_\omega) = \tau$ P-almost everywhere. This shows that for P-almost all ω, A_ω has measure τ. So for P-almost all ω, $|D(X^{(n)}(s),s) - D^{(n)}(X^{(n)}(s),s)|$ converges to 0 for λ-almost all s. Moreover, because of expression (13) for D and expression (15) for $D^{(n)}$ and Assumptions 7.3(b) and 7.1(c), $e^{C\cdot s}|D(\cdot,s) - D^{(n)}(\cdot,s)|$ is bounded by $2e^{C\cdot\tau}C_2/\varepsilon$ on Ω'. Therefore, for almost all ω Lebesgue's dominated convergence theorem can be applied on the integral of $e^{C\cdot s}|D(X^{(n)}(s),s) - D^{(n)}(X^{(n)}(s),s)|$ with respect to λ, $\int_{[0,\tau]} e^{C\cdot s}|D(X^{(n)}(s),s) - D^{(n)}(X^{(n)}(s),s)|\, ds$, implying that for almost all ω this integral converges to 0 as $n \to \infty$. With equation (19), this implies that

$$\sup_{t\in[0,\tau]} \left|X^{(n)}(t) - X(t)\right| \to 0 \qquad \text{a.s.} \tag{22}$$

Since the almost sure limit of a sequence of random variables is measurable if the σ-algebra is complete, measurability of $X(t)$ follows immediately from measurability of the $X^{(n)}$.

7.12. *Conclusion.* This section shows that since $X^{(n)}(t) \sim Y^{(t)}$ given $\overline{Z}_t^{(n)}$ (see Section 7.8) and $X^{(n)}(t) \to X(t)$ a.s. (see Section 7.11), $X(t) \sim Y^{(t)}$ given $\overline{Z}_t$. This completes the proof.

It is well known (see, e.g., [28]; Lemma D.10 provides a formal proof for this conditional version) that $X(t) \sim Y^{(t)}$ given $\overline{Z}_t$ if

$$E\big[f\big(X(t)\big)|\overline{Z}_t\big] - E\big[f\big(Y^{(t)}\big)|\overline{Z}_t\big] = 0 \qquad \text{a.s.}$$

for every bounded Lipschitz continuous function $f : \mathbb{R} \to \mathbb{R}$. Suppose without loss of generality that f is bounded by 1 and has Lipschitz constant L. Then, using the triangle inequality,

$$\begin{aligned}
&\big|E\big[f\big(X(t)\big)|\overline{Z}_t\big] - E\big[f\big(Y^{(t)}\big)|\overline{Z}_t\big]\big| \\
&\quad\le \big|E\big[f\big(X(t)\big)|\overline{Z}_t\big] - E\big[f\big(X(t)\big)|\overline{Z}_t^{(n)}\big]\big| \\
&\qquad + \big|E\big[f\big(X(t)\big)|\overline{Z}_t^{(n)}\big] - E\big[f\big(X^{(n)}(t)\big)|\overline{Z}_t^{(n)}\big]\big| \\
&\qquad + \big|E\big[f\big(X^{(n)}(t)\big)|\overline{Z}_t^{(n)}\big] - E\big[f\big(Y^{(t)}\big)|\overline{Z}_t\big]\big|.
\end{aligned}$$

Because of Jensen's inequality, the second term is bounded by $E[|f(X(t)) - f(X^{(n)}(t))||\overline{Z}_t^{(n)}]$, which is bounded by $E[L|X(t) - X^{(n)}(t)| \wedge 2|\overline{Z}_t^{(n)}]$ since f is Lipschitz continuous with Lipschitz constant L and bounded by 1. Because $X^{(n)}(t) \sim Y^{(t)}$ given $\overline{Z}_t^{(n)}$, the third term is equal to $|E[f(Y^{(t)})|\overline{Z}_t^{(n)}] - E[f(Y^{(t)})|\overline{Z}_t]|$. Therefore,

$$\begin{aligned}
&\big|E\big[f\big(X(t)\big)|\overline{Z}_t\big] - E\big[f\big(Y^{(t)}\big)|\overline{Z}_t\big]\big| \\
&\quad\le \big|E\big[f\big(X(t)\big)|\overline{Z}_t\big] - E\big[f\big(X(t)\big)|\overline{Z}_t^{(n)}\big]\big| \\
&\qquad + E\big[L|X(t) - X^{(n)}(t)| \wedge 2|\overline{Z}_t^{(n)}\big] \\
&\qquad + \big|E\big[f\big(Y^{(t)}\big)|\overline{Z}_t^{(n)}\big] - E\big[f\big(Y^{(t)}\big)|\overline{Z}_t\big]\big| \qquad \text{a.s.}
\end{aligned} \tag{23}$$

We show that the right-hand side converges in probability to zero. On the first and the last term, Lévy's upward theorem (see, e.g., [29] p. 134) can be applied, since the integrands are bounded by 1. Lévy's upward theorem leads to

$$E\big[f\big(X(t)\big)|\overline{Z}_t^{(n)}\big] \to E\left[f\big(X(t)\big)\middle|\sigma\left(\bigcup_{n=1}^{\infty}\overline{Z}_t^{(n)}\right)\right]$$

and

$$E\big[f\big(Y^{(t)}\big)|\overline{Z}_t^{(n)}\big] \to E\left[f\big(Y^{(t)}\big)\middle|\sigma\left(\bigcup_{n=1}^{\infty}\overline{Z}_t^{(n)}\right)\right]$$

as $n \to \infty$. Thus, with Lemma A.2 in the Appendix, both the first and the last term of equation (23) converge to 0 almost surely. The second term converges to 0 in probability since it is almost surely nonnegative and its expectation converges to 0:

$$E\big(E\big[L\big|X(t) - X^{(n)}(t)\big| \wedge 2\big|\overline{Z}_t^{(n)}\big]\big) = E\big(L\big|X(t) - X^{(n)}(t)\big| \wedge 2\big) \to 0$$

because of Lebesgue's dominated convergence theorem and the fact that $X^{(n)}(t)$ converges almost surely to $X(t)$.

Thus, $|E[f(X(t))|\overline{Z}_t] - E[f(Y^{(t)})|\overline{Z}_t]|$ is bounded by a random variable which converges in probability to 0. Hence, this first random variable is almost surely equal to 0. Therefore, indeed $X(t)$ mimics $Y^{(t)}$ in the sense that $X(t)$ has the same distribution as $Y^{(t)}$ given $\overline{Z}_t$.

7.13. *Mimicking counterfactual outcomes*: *Discrete-continuous time*. In certain situations, there are specific times t with $P(t \text{ is a jump time of } Z) > 0$. For finitely many such times t, the proof in Section 7 can be adapted by adding these finitely many times to the grid, for each n.

8. Mimicking counterfactual survival outcomes.

8.1. *Introduction*. This section indicates how to prove that $X(t)$ mimics $Y^{(t)}$ in the sense that $X(t)$ has the same distribution as $Y^{(t)}$ given the covariate and treatment history $\overline{Z}_t$, under conditions aimed at survival. The conditions are similar to the ones in Section 7, but adapted to survival as the outcome of interest. The proof also follows roughly the same lines as the one for other outcomes, but some changes are necessary. A full proof can be found in Lok [10] Section B.

If covariates and treatment were measured at time t, it cannot be avoided to include in $\overline{Z}_t$ whether or not a person was alive at time t: what are a person's covariates if he or she is dead? Therefore, we include in $Z(t)$ an indicator for whether or not a person is alive at time t. Thus, if a person died at or before time t, the survival time can be read from $\overline{Z}_t$.

The conditions in Section 7 usually exclude survival as the outcome of interest, since if the outcome is survival the support condition 7.1, saying that all $F_{Y^{(t+h)}|\overline{Z}_t}$ have the same bounded support $[y_1, y_2]$, will not hold: $\overline{Z}_t$ includes the covariate-measurements and treatment until time t, and given that a person is dead at time t and given his or her survival time, the distribution of this survival time cannot have the fixed support $[y_1, y_2]$, independent of t. Also given that a person is alive at time t, the survival time often does not have the fixed support $[y_1, y_2]$: one often expects that t is the left limit of the support, and obviously the left limit of the support should be greater than or equal to t.

I make two extra assumptions. The first is a straightforward consistency assumption, stating that stopping treatment after death does not change the survival time. The second extra assumption states that there is no instantaneous effect of

treatment at the time the person died (notice that the difference between $Y^{(Y)}$, the outcome with treatment stopped at the survival time Y, and Y is in treatment *at* time Y).

ASSUMPTION 8.1 (consistency). $Y^{(t)} = Y$ on $\{\omega : Y \le t\} \cup \{\omega : Y^{(t)} \le t\}$.

ASSUMPTION 8.2 (no instantaneous effect of treatment at the time the person died). $Y^{(t)} = Y$ on $\{\omega : Y = t\} \cup \{\omega : Y^{(t)} = t\}$.

As can be expected, these assumptions imply that treatment in the future does not cause or prevent death at present, see Lok [10] Section B.

For survival outcomes, this article uses the following minor adaptation of the definition of D,

$$D(y,t;\overline{Z}_t) = \begin{cases} 0 \quad \text{if } \overline{Z}_t \text{ indicates the person is dead at } t \text{ or } y < t \\ \frac{\partial}{\partial h}|_{h=0}\big(F^{-1}_{Y^{(t+h)}|\overline{Z}_t} \circ F_{Y^{(t)}|\overline{Z}_t}\big)(y) \quad \text{otherwise, for } y > t \\ \lim_{y \downarrow t} D(y,t;\overline{Z}_t) \quad \text{otherwise, for } y = t, \end{cases} \tag{24}$$

as we explain now. First, remark that considering the interpretation of $D(y,t;\overline{Z}_t)$ as the infinitesimal effect of a short duration of treatment directly after t on survival, $D(y,t;\overline{Z}_t)$ should be zero if $\overline{Z}_t$ indicates the person is dead at time t. Although in that case indeed $F_{Y^{(t+h)}|\overline{Z}_t}$ and $F_{Y^{(t)}|\overline{Z}_t}$ are almost surely the same for every $h \ge 0$, since withholding treatment after death does not change the survival time, $F^{-1}_{Y^{(t+h)}|\overline{Z}_t}$ will often not exist. Therefore, if $\overline{Z}_t$ indicates the person is dead at time t, this article just formally defines $D(y,t;\overline{Z}_t)$ to be zero. Next, consider $y < t$. Notice that considering the interpretation of $D(y,t;\overline{Z}_t)$ as the infinitesimal effect of treatment directly after time t on the survival-quantile y, $D(y,t;\overline{Z}_t)$ should be zero for $y < t$ since treatment at or after time t should not cause or prevent death at or before time t, so it should not affect quantiles of the survival curve before time t. Indeed if $\overline{Z}_t$ indicates that the person is alive at time t, $F_{Y^{(t+h)}|\overline{Z}_t}(y) = F_{Y|\overline{Z}_t}(y) = 0$ for $y \le t$ for all $h \ge 0$, but also for these y, $F^{-1}_{Y^{(t+h)}|\overline{Z}_t}(y)$ often does not exist. Therefore, this article defines $D(y,t;\overline{Z}_t) = 0$ for $y < t$. In order to make D continuous on $y \ge t$ in between the jump times of Z, we define $D(t,t;\overline{Z}_t) = \lim_{y \downarrow t} D(y,t;\overline{Z}_t)$. This limit exists under the conditions in Section 8.2. It is not necessarily equal to zero.

Notice that the area where D is possibly nonzero is $(y,t) \in [0,\infty) \times [0, \min\{Y,\tau\}] : y \ge t$. Therefore, if $Y < \tau$, the solution to the differential equation $X(t)$ is equal to Y for $t \in [Y,\tau]$. An example of such $X(t)$ is shown in Figure 1 (right).

In the case of a survival outcome, right censoring is common. For right censoring, [22] proposed the artificial censoring estimator. A slight adaptation of this estimator is presented in Section 8.4.

8.2. *Mimicking counterfactual survival outcomes*: *Assumptions and result*. This section presents precise conditions under which $X(t)$ mimics $Y^{(t)}$, for survival outcomes, following Section 7.2.1 (Lok [10] Section B provides conditions similar to Section 7.2). We choose versions of $F_{Y^{(t+h)}|\overline{Z}_t}$ that (a) are consistent with the fact that treatment after death is irrelevant, and (b) satisfy all regularity conditions below. These versions are used in the definition of D for survival outcomes, and everywhere in the proof.

ASSUMPTION 8.3 (Regularity conditions).

- *(support)*. There exists a finite number $y_2 \geq \tau$ such that:

  (a) If $Y > t$, all $F_{Y^{(t+h)}|\overline{Z}_t}$, for $h \geq 0$ and $t \in [0, \tau]$, have support $[t, y_2]$.
  (b) If $Y > t$, all $F_{Y^{(t+h)}|\overline{Z}_t}$, for $h \geq 0$ and $t \in [0, \tau]$, have a continuous nonzero density $f_{Y^{(t+h)}|\overline{Z}_t}(y)$ on $y \in [t + h, y_2]$.
  (c) There exists an $\varepsilon > 0$ such that for all $\omega \in \Omega$ and t with $Y > t$, $f_{Y^{(t)}|\overline{Z}_t}(y) > \varepsilon$ for $y \in [t, y_2]$.

- *(smoothness)*. For every $\omega \in \Omega$

  (a) If Z does not jump in (t_1, t_2) and $Y > t_1$, the restriction of $(y, t, h) \to F_{Y^{(t+h)}|\overline{Z}_t}(y)$ to $\{(y, t, h) \in [t_1, y_2] \times [t_1, t_2) \times \mathbb{R}_{\geq 0} : y \geq t + h\}$ is C^1 in (y, t, h).
  (b) The derivatives of $F_{Y^{(t+h)}|\overline{Z}_t}(y)$ $(y > t + h)$ with respect to y and h are bounded by constants C_1 and C_2, respectively.
  (c) $\frac{\partial}{\partial y} F_{Y^{(t)}|\overline{Z}_t}(y)$ and $\frac{\partial}{\partial h}|_{h=0} F_{Y^{(t+h)}|\overline{Z}_t}(y)$ $(y > t)$ have derivatives with respect to y which are bounded by constants L_1 and L_2, respectively.
  (d) For all $\omega \in \Omega$ and t with $Y > t$, $F_{Y|\overline{Z}_t}(y)$ is continuous and strictly increasing on its support $[t, y_2]$.

THEOREM 8.4. *Suppose that Regularity Condition 8.3 is satisfied. Then* $D(y, t; \overline{Z}_t)$ *as defined in equation* (24) *exists. Furthermore, for every* $\omega \in \Omega$ *there exists exactly one continuous solution* $X(t)$ *to* $dX(t)/dt = D(X(t), t; \overline{Z}_t)$ *with final condition* $X(\tau) = Y$. *If also Assumptions 4.1, 8.1 and 8.2* (*consistency and no instantaneous treatment effect at time of death*) *are satisfied then this* $X(t)$ *has the same distribution as* $Y^{(t)}$ *given* $\overline{Z}_t$ *for all* $t \in [0, \tau]$.

8.3. *Outline of the proof.* The proof of Theorem 8.4 follows the same lines as the proof of Theorem 4.2. The one essential difference between survival outcomes and nonsurvival outcomes is: if $\overline{Z}_t$ indicates the person is alive at time t, $X(t)$ should be greater than t, since we want $X(t)$ to have the same distribution as $Y^{(t)}$ given $\overline{Z}_t$ ($Y^{(t)} > t$ in that case because of consistency Assumption 8.1). This leads to an additional problem in the proof for the continuous-time case, namely: how to prove that the solution stays above the line $y = t$ for $t \in [0, Y]$? I solve this additional problem in Lok [10] Section B by showing that, under the assumptions of Section 8.2, $D(t, t; \overline{Z}_t) \leq 1$. In addition, extra technical problems arise because

the smoothness conditions have to be adapted to the survival setting; see Lok [10] Section B for details.

8.4. *Survival outcomes and right censoring.* In the case of a survival outcome, right censoring is common. Robins [22] proposed the artificial censoring estimator for administrative censoring. That is censoring due to end-of-follow-up because the study ends. The idea behind artificial censoring is that, instead of adding $X(t)$ or $X(0)$ to the model for predicting treatment changes (see Theorem 5.2), one could add a function $\tilde{X}(0)$ of $X(0)$ and the censoring time C, which is observed for all patients. The artificial censoring estimator treats the censoring time C as a baseline covariate. This is justified in the case of censoring due to study closure, because in this case C only depends on the date a patient enrolled in the study. Conditional on the value of $\overline{Z}_{t-}$, functions of $X(t)$ and $\overline{Z}_{t-}$ are not predictive of treatment changes (Theorem 5.2). Therefore, conditional on $\overline{Z}_{t-}$, $\tilde{X}(0)$ is not predictive of treatment changes either. This produces an estimation procedure for ψ analogous to that in Theorem 5.2, but that allows for right censoring.

We slightly adapt this procedure, and propose to add a function of $X(t)$ and C to the model for the prediction of treatment changes. In particular, for D as in equation (3) and for $\min(Y, C) \geq t$, we propose to add to the prediction model of treatment changes the function $\tilde{X}(t, \psi) = \min(X_\psi(t), C(t, \psi))$, with

$$C(t, \psi) = \begin{cases} C & \text{if } \psi \geq 0 \\ t + e^{\psi}(C - t) & \text{if } \psi < 0. \end{cases}$$

As required, $\tilde{X}(t, \psi)$ is a function of $X_\psi(t)$ and C. In addition, we will show that both for the case that $\psi \geq 0$ and for the case that $\psi < 0$, $\tilde{X}(t, \psi)$ is observed for all patients. This follows from the fact that

$$\begin{aligned} \tilde{X}(t, \psi) &= \min\bigl(X^*(t, \psi), C(t, \psi)\bigr), \\ &\text{with } X^*(t, \psi) = t + \textstyle\int_t^{\min(Y,C)} e^{\psi 1_{\text{no prophylaxis at } s}}\, ds, \end{aligned} \tag{25}$$

which is observed for all patients. For $\psi \geq 0$, equation (25) follows from

$$\begin{aligned} \tilde{X}(t, \psi) &= \min\left(t + \int_t^Y e^{\psi 1_{\text{no prophylaxis at } s}}\, ds, C\right) \\ &= \min\left(t + \int_t^Y e^{\psi 1_{\text{no prophylaxis at } s}}\, ds,\right. \\ &\qquad \left. t + \int_t^C e^{\psi 1_{\text{no prophylaxis at } s}}\, ds, C\right) \\ &= \min\bigl(X^*(t, \psi), C(t, \psi)\bigr), \end{aligned}$$

where for the second equality we used that for $\psi \geq 0$, $t + \int_t^C e^{\psi 1_{\text{no prophylaxis at } s}} ds \geq C$. For $\psi < 0$, equation (25) follows from

$$\begin{aligned}\tilde{X}(t,\psi) &= \min\Big(t + \int_t^Y e^{\psi 1_{\text{no prophylaxis at } s}} ds, t + e^{\psi}(C-t)\Big) \\ &= \min\Big(t + \int_t^Y e^{\psi 1_{\text{no prophylaxis at } s}} ds, \\ &\qquad t + \int_t^C e^{\psi 1_{\text{no prophylaxis at } s}} ds, t + e^{\psi}(C-t)\Big) \\ &= \min\big(X^*(t,\psi), C(t,\psi)\big),\end{aligned}$$

where for the second equality we used that for $\psi < 0$, $t + \int_t^C e^{\psi 1_{\text{no prophylaxis at } s}} ds \geq t + e^{\psi}(C-t)$.

For the case that $\psi < 0$, some patients are "artificially" censored, since if $C > t$, $C(t,\psi) = t + e^{\psi}(C-t) < C$. Artificial censoring produces a subclass of the estimators considered in Theorem 5.2 allowing h_t to depend on ψ: $h_{t,\psi}(X_\psi(t), \overline{Z}_{t-}) = 1_{\min(Y,C)\geq t}\tilde{h}_t(\min(X_\psi(t), C(t,\psi)), \overline{Z}_{t-})$ (notice that $1_{\min(Y,C)\geq t}$ is a function of $\overline{Z}_{t-}$). In general, one could add to the prediction model for treatment changes any function of $X_\psi(t)$ and C that is observed for all patients. Robins [22] suggests to also consider adding $\Delta(t,\psi) = 1_{\tilde{X}(t,\psi)\leq C(t,\psi)}$ to the model for the prediction of treatment changes. Since both $\tilde{X}(t,\psi)$ and $C(t,\psi)$ are observed for all patients, so is $\Delta(t,\psi)$. Thus, the above reasoning shows that this procedure leads to consistent estimation of the treatment effect as well.

The procedure above can easily be adapted to for example model (5), by replacing $C(t,\psi)$ accordingly. To be more specific, for that case one could use

$$C(t,\psi) = t + e^{\min(\psi_1,0)+\min(\psi_2,0)+\min(\psi_3,0)}(C-t).$$

9. Simulation study. In the simulation study, we calibrated the distributions of the variables and the parameter values to HIV/AIDS data, perhaps the most salient example of application of structural nested models in the empirical literature. We focus on the first two years since HIV diagnosis. Time zero is the time of HIV diagnosis. The outcome variable is the CD4 count, a commonly used marker of the state of the immune system of HIV-positive patients. The usual treatment for HIV-positive patients is ART, antiretroviral treatment. ART is not always initiated immediately after diagnosis. ART initiation time often depends on the last measured CD4 count. When the CD4 count is at or below 350 copies/ml, HIV-positive patients are much more likely to initiate ART than when the CD4 count is above 350 copies/ml. Lok [10] Section C describes how we generated the data for the simulation study in detail, including distributions and parameter values. This section provides an overview.

In this simulation study, no one is treated at time zero, and once treatment is initiated, it is never stopped. $Y^{(t)}$ is the counterfactual outcome had treatment been

as given in reality until time t, and continued or initiated after that. For example, if treatment was initiated by time t for a particular patient, $Y^{(t)}$ is the observed outcome for that patient, since he or she was already treated at time t and treatment is never stopped. On the other hand, if treatment was not initiated by time t, $Y^{(t)}$ is the outcome had treatment been initiated at time t. Thus, in the definition of $Y^{(t)}$ in Section 2, the switch at time t to "some kind of baseline treatment regime $\overline{0}$" is, in this case, "treat continuously" from time t onwards. In the simulations, we study a setting with $t \in [0, 2]$. The subscript $_t$ indicates the treatment initiation time, so for example $L_{1,t}$ indicates (counterfactual) covariates at time 1 under "treatment started at time t." Similarly, the subscript $_\infty$ indicates (counterfactual) variables under no treatment. For example, $L_{2,\infty}$ indicates (counterfactual) covariates at time 2 under no treatment. In the simulation design, the counterfactual covariates L are as follows:

$$\begin{aligned}
L_0 &= \tilde{L}_0 + e_0, \\
L_{1,\infty} &= \tilde{L}_0 - \beta_0 + e_{1,\infty}, \\
L_{2,\infty} &= \tilde{L}_0 - 2\beta_0 + e_{2,\infty} \\
L_{1,t} &= \tilde{L}_0 - \beta_0 + \theta(1-t) + e_{1,t} \qquad \text{for } t \in [0, 1], \text{ and } L_{1,\infty} \text{ otherwise} \\
L_{2,t} &= \tilde{L}_0 - 2\beta_0 + \psi(2-t) + e_{2,t},
\end{aligned}$$

where $\tilde{L}_0$ and the $e_{j,t}$ are random variables with values in $\mathbb{R}$. Notice that $(1-t)$ and $(2-t)$ are simply the durations of treatment until the respective covariate measurements. We assume that the $e_{j,t}$ $(j = 0, 1, 2)$ are independent of $\tilde{L}_0$, and that the $e_{2,t}$ have a distribution function which does not depend on t. We also assume that the $e_{2,t}$ are independent of all previous variables (and of the treatment initiation time, T, described below). In the simulations, $\psi \geq 0$ (a similar study could have been done for $\psi < 0$). We define $Y_t = L_{2,t}$, the counterfactual outcome with treatment initiated at time t, which could potentially be observed at time 2.

We show in Lok [10] Section C that the outcome processes adopted in our simulation study are not rank preserving. This is easily seen because with probability one, two patients with the same observed data do not have the same value of $\tilde{L}_0$.

Suppose that the hazard of the treatment initiation time, T, given the covariate history at time t and given that treatment was not initiated before time t, is piecewise constant as follows:

$$\lambda_T(t) = \begin{cases} \lambda_0^{(0)} & \text{if } L_0 > c_0 \text{ and } t \in [0, 1] \\ \lambda_1^{(0)} & \text{if } L_0 \leq c_0 \text{ and } t \in [0, 1] \\ \lambda_0^{(1)} & \text{if } L_{1,\infty} > c_1 \text{ and } t \in (1, 2] \\ \lambda_1^{(1)} & \text{if } L_{1,\infty} \leq c_1 \text{ and } t \in (1, 2], \end{cases}$$

for constants c_0 and c_1 in $\mathbb{R}$. Notice that T depends on $\tilde{L}_0$, $e_{0,\infty}$ and $e_{1,\infty}$, if $\lambda_0^{(0)} \neq \lambda_1^{(0)}$ or $\lambda_0^{(1)} \neq \lambda_1^{(1)}$.

In the simulation study, treatment can be initiated in continuous time, but the covariates are only measured at times 0, 1 and 2, so that the treatment and covariate history up to time t, $\overline{Z}_t$, consists of the treatment information up to time t and L_0, (L_0, L_1) or (L_0, L_1, L_2), depending on whether $t \in [0, 1)$, $t \in [1, 2)$ or $t = 2$. In the simulations, treatment affects later outcomes, and time-dependent covariates (L_1) which depend on previous treatment also predict future treatment and the outcome of interest. This is the type of setting structural nested models were developed for.

Lok [10] Section C shows that for this data generating mechanism,

$$D(y, t; \overline{Z}_t) = -\psi 1_{\text{untreated at } t}.$$

Then it follows from the definition of X_ψ that

$$X_\psi(t) = Y + \psi\big(\min(T, 2) - t\big)1_{T>t},$$

where $(\min(T, 2) - t)1_{T>t}$ is the duration of the patient not being on treatment between time t and time 2.

As shown in Lok [10] Section C, a consistent estimator of ψ can be defined as follows. In the first step, the nuisance parameters $(\lambda_0^{(0)}, \lambda_1^{(0)}, \lambda_0^{(1)}, \lambda_1^{(1)})$ are estimated using maximum likelihood theory. In the second step, ψ is estimated as $\hat{\psi} = -\sum_{i=1}^n A_{1i} / \sum_{i=1}^n A_{2i}$, where

$$\begin{aligned}
A_{1i} = &-Y_i\big(Z_i(0)\hat{\lambda}_1^{(0)} + \big(1 - Z_i(0)\big)\hat{\lambda}_0^{(0)}\big)\min(T_i, 1) \\
&- Y_i\big(Z_i(1)\hat{\lambda}_1^{(1)} + \big(1 - Z_i(1)\big)\hat{\lambda}_0^{(1)}\big)\big(1 - \delta_i^{(0)}\big)\big(\min(T_i, 2) - 1\big) \\
&+ Y_i\delta_i^{(0)} + Y_i\delta_i^{(1)}, \\
A_{2i} = &-\big(Z_i(0)\hat{\lambda}_1^{(0)} + \big(1 - Z_i(0)\big)\hat{\lambda}_0^{(0)}\big)\min(T_i, 1)\min(T_i, 2) \\
&- \big(Z_i(1)\hat{\lambda}_1^{(1)} + \big(1 - Z_i(1)\big)\hat{\lambda}_0^{(1)}\big)\big(1 - \delta_i^{(0)}\big)\big(\min(T_i, 2) - 1\big)^2 \\
&+ \delta_i^{(0)}\min(T_i, 2) + \delta_i^{(1)}\big(\min(T_i, 2) - 1\big),
\end{aligned}$$

$\delta_i^{(0)} = 1_{T_i \le 1}$, $\delta_i^{(1)} = 1_{1 < T_i \le 2}$, $Z_i(0) = 1_{L_0 \le c_0}$, and $Z_i(1) = 1_{L_1 \le c_1}$.

We ran a simulation study with $n = 500$, 1000, 2000, 5000 and 10,000, with 5000 repetitions each. The results are presented in Table 1. As detailed in Lok [10] Section C, setting 1 has the least noise around the signals, and setting 3 the most.

In this simulation study, both for small and large samples, the bias of the estimators is small. In all three settings and for all sample sizes considered (including the small sample size $n = 100$), the MSE of the estimators arises mostly from the variance, not from the bias. Also, if the true parameter ψ equals 300 as in this simulation study, for $n = 500$, $\sqrt{MSE}/\psi = 0.04$ in setting 1 and 0.08 in setting 3. Thus, the estimates are already precise in relatively small samples. Because, as shown in Lok [10] the MSE in this simulation study does not depend on the true parameter, ψ, a larger sample size would be required to obtain precise estima-

TABLE 1
Simulations. Mean Squared Errors (MSE) and bias. 5000 repetitions each

	Setting 1			Setting 2			Setting 3		
n	MSE	$\frac{\text{MSE}\times n}{1000}$	bias	MSE	$\frac{\text{MSE}\times n}{1000}$	bias	MSE	$\frac{\text{MSE}\times n}{1000}$	bias
100	747	75	−0.28	1907	191	−0.040	2875	287	−0.39
500	146	73	−0.22	356	178	−0.29	542	271	−0.61
1000	72	72	−0.10	176	176	−0.068	268	268	−0.23
2000	35	70	−0.11	89	179	0.051	138	275	−0.08
5000	14	69	−0.067	35	175	−0.0040	54	268	−0.05
10,000	6.6	66	−0.066	18	178	−0.022	27	270	−0.06

tors of small true parameter values ψ. We conclude that in this simulation study, continuous-time structural nested models perform extremely well.

10. Discussion. Structural nested models have become a major part of statistical tools for estimation of the effect of time varying treatments, in the presence of time-dependent confounding by indication; see, for example, [30] for a discrete-time application and, for example, [15, 19, 22, 24, 25, 27] and [8, 9] for continuous-time applications. Structural nested models in continuous time are useful to estimate the effect of a treatment that can be initiated at any point in time, and for which a short duration of treatment has a small effect on the outcome of interest. In contrast with discrete-time structural nested models, in the case of survival outcomes, the resulting parameter estimates can often be interpreted as rates. So far, continuous-time analyses relied on (local) rank preservation. The main result of the current article is to prove that for continuous-time structural nested models, assumptions about the joint distributions of counterfactuals or deterministic treatment effects/ (local) rank preservation are not necessary to "mimic counterfactual outcomes," and, based on that, to consistently estimate treatment effects. This article provides a proof for outcomes that are measured at the end of the study as well as a proof for survival outcomes. Important public health decisions are based on analyses with continuous-time structural nested models, so it is important to relax unverifiable and disputable assumptions underlying these analyses.

An interesting topic for future research is to investigate whether the support conditions 7.1 or 8.3 can be weakened, for example, to an assumption about the support varying in a differentiable way between the jump times of the covariate and treatment process Z. We expect that in that case one has to assume that where Z jumps, the support of $Y^{(t)}$ given $\overline{Z}_t$ gets smaller or stays the same as t increases (see Figure 2). Otherwise, $X(t)$ may move out of the support of $Y^{(t)}$ given $\overline{Z}_t$ (recall that X is the solution to a differential equation with *final* condition). It is reasonable to assume that the support of $Y^{(t)}$ given $\overline{Z}_t$ gets smaller or stays the same as t increases, since more information about $\overline{Z}$ should not enlarge the range of $Y^{(t)}$.

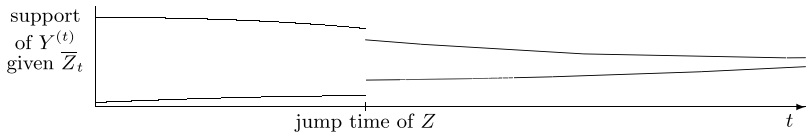

FIG. 2. *Example of support of* $Y^{(t)}$ *given* $\overline{Z}_t$.

A problem which may occur without a support condition is that the denominator in equation (13) (the quotient expression for D) or in equations (14) or (15) (the quotient expression for $D^{(n)}$) may tend to 0, which may "blow up" D or $D^{(n)}$. In that case, it might help to assume that there exists a constant C such that (a) for all $\omega \in \Omega$, t and y, $F^{-1}_{Y^{(t+h)}|\overline{Z}_t} \circ F_{Y^{(t)}|\overline{Z}_t}(y) - y \le C \cdot h$, and (b) for all t, y and $B \subset \overline{Z}_t$ with $P(\overline{Z}_t \in B) > 0$, $F^{-1}_{Y^{(t+h)}|\overline{Z}_t \in B} \circ F_{Y^{(t)}|\overline{Z}_t \in B}(y) - y \le C \cdot h$. This assumption does not look unreasonable if there is no "instantaneous treatment effect." It is to be expected that under this assumption both D and $D^{(n)}$ are bounded by C.

Based on the results of the current article, [11] shows that also if a semiparametric Cox model is used to predict treatment changes in Theorem 5.2, the resulting estimating equations for the treatment effect are unbiased. However, the estimating equations are no longer of the form of an average of terms that are independent for the different persons. Thus, consistency and asymptotic normality for this situation constitute interesting topics for future research.

APPENDIX A

This Appendix contains results that are frequently used in the main article. Lok [10] is more elaborate. The first theorem is a corollary of a theorem in [4] Chapter 2, see Lok [10] Section G.

THEOREM A.1. *Suppose that* I *is a closed interval in* $\mathbb{R}$, $f : I \times [y_1, y_2] \to \mathbb{R}$ *is continuous with for all* $t \in I$, $f(t, y_1) = f(t, y_2) = 0$ *and* $C : I \to [0, \infty)$ *is continuous, and suppose that*

$$|f(t, y) - f(t, z)| \le C(t)|y - z| \tag{26}$$

for all $t \in I$ *and* $y, z \in [y_1, y_2]$. *Then, for every* $t_0 \in I$ *and* $y_0 \in [y_1, y_2]$, *there exists a unique solution* $y(t)$ *of* $y'(t) = f(t, y(t))$ *with* $y(t_0) = y_0$, *and this solution is defined for all* $t \in I$. *Furthermore,* $y(t) \in [y_1, y_2]$ *for all* $t \in I$. *Suppose that* $g : I \times [y_1, y_2] \to \mathbb{R}$ *is continuous and* $z : I \to [y_1, y_2]$ *is a solution of* $z'(t) = g(t, z(t))$. *Then*

$$\begin{aligned} |y(t) - z(t)| &\le e^{\int_t^{t_0} C(s)\,ds} |y(t_0) - z(t_0)| \\ &\quad + \int_t^{t_0} e^{\int_t^s C(\eta)\,d\eta} |f(t, z(t)) - g(s, z(s))|\,ds \end{aligned}$$

for all $t, t_0 \in I$ *with* $t \le t_0$.

The proof of the following lemma can be found in Lok [10] Section D.

LEMMA A.2. *Let X be a random variable with $E|X| < \infty$, and let $\overline{Z}_t$ be a random variable with values in $\overline{\mathcal{Z}}_t$, the space of cadlag functions on $[0, t]$ provided with the projection σ-algebra, with $P(Z \text{ jumps at } t) = 0$. Then any version of $E[X|\sigma(\bigcup_{n=1}^{\infty} \overline{Z}_t^{(n)})]$, with $\overline{Z}_t^{(n)}$ as defined in Section 7.6, is also a version of $E[X|\overline{Z}_t]$.*

Acknowledgments. I am indebted to Richard Gill and Aad van der Vaart for their support, insight and encouragement on this project. I also thank James Robins for fruitful discussions, and Nell Sedransk for constructive comments on the writing. I thank Susan Little for allowing me to use the AIEDRP data to calibrate the distributions in the simulation study.

SUPPLEMENTARY MATERIAL

Web-Appendix with "Mimicking counterfactual outcomes to estimate causal effects" (DOI: 10.1214/15-AOS1433SUPP; .pdf). This Web-Appendix provides mathematical details about mimicking counterfactual survival outcomes, additional information on the simulation study, and theorems used in the main text.

REFERENCES

[1] BAUER, H. (1972). *Probability Theory and Elements of Measure Theory*. Holt, Rinehart & Winston, New York.
[2] BILLINGSLEY, P. (1968). *Convergence of Probability Measures*. Wiley, New York.
[3] COX, D. R. and OAKES, D. (1984). *Analysis of Survival Data*. Chapman & Hall, London. MR0751780
[4] DUISTERMAAT, J. J. and ECKHAUS, W. (1995). *Analyse van Gewone Differentiaalvergelijkingen. Epsilon Uitgaven*. Epsilon, Utrecht.
[5] GILL, R. D. and ROBINS, J. M. (2001). Causal inference for complex longitudinal data: The continuous case. *Ann. Statist.* **29** 1785–1811. MR1891746
[6] HERNÁN, M. A., BRUMBACK, B. and ROBINS, J. M. (2000). Marginal structural models to estimate the causal effect of zidovudine on the survival of HIV-positive men. *Epidemiology* **11** 561–570.
[7] HOLLAND, P. W. (1986). Statistics and causal inference. *J. Amer. Statist. Assoc.* **81** 945–970. MR0867618
[8] KEIDING, N. (1999). Event history analysis and inference from observational epidemiology. *Stat. Med.* **18** 2353–2363.
[9] KEIDING, N., FILIBERTI, M., ESBJERG, S., ROBINS, J. M. and JACOBSEN, N. (1999). The graft versus leukemia effect after bone marrow transplantation: A case study using structural nested failure time models. *Biometrics* **55** 23–28.
[10] LOK, J. J (2017). Supplement to "Mimicking counterfactual outcomes to estimate causal effects." DOI:10.1214/15-AOS1433SUPP.
[11] LOK, J. J. (2001). Statistical modelling of causal effects in time, Ph.D. thesis, Department of Mathematics, Free Univ. Amsterdam. http://www.math.vu.nl/research/theses/pdf/lok.pdf.
[12] LOK, J. J. (2007). Structural nested models and standard software: A mathematical foundation through partial likelihood. *Scand. J. Stat.* **34** 186–206. MR2325250

[13] LOK, J. J. (2008). Statistical modelling of causal effects in continuous time. *Ann. Statist.* **36** 1464–1507.

[14] LOK, J. J., GILL, R. D., VAN DER VAART, A. W. and ROBINS, J. M. (2004). Estimating the causal effect of a time-varying treatment on time-to-event using structural nested failure time models. *Stat. Neerl.* **58** 271–295.

[15] MARK, S. D. and ROBINS, J. M. (1993). Estimating the causal effect of smoking cessation in the presence of confounding factors using a rank preserving structural failure time model. *Stat. Med.* **12** 1605–1628.

[16] POLLARD, D. (2001). *A User's Guide to Measure Theoretic Probability*. Cambridge Univ. Press, Cambridge.

[17] ROBINS, J. M. (1987). A graphical approach to the identification and estimation of causal parameters in mortality studies with sustained exposure periods. *Journal of Chronic Disease* **40** 139S–161S.

[18] ROBINS, J. M. (1989). The analysis of randomized and non-randomized AIDS treatment trials using a new approach to causal inference in longitudinal studies. In *Health Service Research Methodology*: *A Focus on AIDS* (L. Sechrest, H. Freeman and A. Bailey, eds.) 113–159. NCHSR, U.S. Public Health Service, Washington, D.C.

[19] ROBINS, J. M. (1992). Estimation of the time-dependent accelerated failure time model in the presence of confounding factors. *Biometrika* **79** 321–334.

[20] ROBINS, J. M. (1995). Causal inference from complex longitudinal data. In *Design and Analysis of Follow-up Studies*. The Netherlands Institute for Health Sciences, Rotterdam. Lecture notes.

[21] ROBINS, J. M. (1997). Causal inference from complex longitudinal data. In *Latent Variable Modeling and Applications to Causality* (*Los Angeles*, *CA*, 1994). *Lect. Notes Stat.* **120** 69–117. Springer, New York. MR1601279

[22] ROBINS, J. M. (1998). Structural nested failure time models. In *Survival Analysis* (P. Armitage and T. Colton, eds.). *Encyclopedia of Biostatistics* **6** 4372–4389. Wiley, Chichester, UK. Section Eds: P. K. Andersen and N. Keiding.

[23] ROBINS, J. M. (1999). Marginal structural models versus structural nested models as tools for causal inference. In *Statistical Models in Epidemiology*: *The Environment and Clinical Trials* (M. E. Halloran and D. Berry, eds.) **116** 95–134. Springer, New York.

[24] ROBINS, J. M., BLEVINS, D., RITTER, G. and WULFSOHN, M. (1992). G-estimation of the effect of prophylaxis therapy for pneumocystis carinii pneumonia on the survival of AIDS patients. *Epidemiology* **3** 319–336.

[25] ROBINS, J. M. and GREENLAND, S. (1994). Adjusting for differential rates of PCP prophylaxis in high- versus low-dose AZT treatment arms in an AIDS randomized trial. *J. Amer. Statist. Assoc.* **89** 737–749.

[26] ROBINS, J. M., HERNÁN, M. A. and BRUMBACK, B. (2000). Marginal structural models and causal inference in epidemiology. *Epidemiology* **11** 550–560.

[27] TILLING, K., STERNE, J. A. and SZKLO, M. (2002). Estimating the effect of cardiovascular risk factors on all-cause mortality and incidence of coronary heart disease using G-estimation: The atherosclerosis risk in communities study. *Am. J. Epidemiol.* **155** 710–718.

[28] VAN DER VAART, A. W. (1998). *Asymptotic Statistics*. *Cambridge Series in Statistical and Probabilistic Mathematics* **3**. Cambridge Univ. Press, Cambridge. MR1652247

[29] WILLIAMS, D. (1991). *Probability with Martingales*. Cambridge Univ. Press, Cambridge.

[30] WITTEMAN, J. C. M., D'AGOSTINO, R. B., STIJNEN, T., KANNEL, W. B., COBB, J. C., DE RIDDER, M. A. J., HOFMAN, A. and ROBINS, J. M. (1998). G-estimation of causal effects: Isolated systolic hypertension and cardiovascular death in the Framingham heart study. *American Journal of Epidemiology* **148** 390–401.

DEPARTMENT OF BIOSTATISTICS
HARVARD SCHOOL OF PUBLIC HEALTH
655 HUNTINGTON AVENUE, BUILDING 2, ROOM 409
BOSTON, MASSACHUSETTS 02115
USA
E-MAIL: jlok@hsph.harvard.edu
URL: http://www.hsph.harvard.edu/judith-lok/