The Annals of Statistics
2008, Vol. 36, No. 3, 1464–1507
DOI: 10.1214/009053607000000820

STATISTICAL MODELING OF CAUSAL EFFECTS IN CONTINUOUS TIME[1]

BY JUDITH J. LOK

Free University of Amsterdam and
Harvard School of Public Health

This article studies the estimation of the causal effect of a time-varying treatment on time-to-an-event or on some other continuously distributed outcome. The paper applies to the situation where treatment is repeatedly adapted to time-dependent patient characteristics. The treatment effect cannot be estimated by simply conditioning on these time-dependent patient characteristics, as they may themselves be indications of the treatment effect. This time-dependent confounding is common in observational studies. Robins [(1992) *Biometrika* **79** 321–334, (1998b) *Encyclopedia of Biostatistics* **6** 4372–4389] has proposed the so-called structural nested models to estimate treatment effects in the presence of time-dependent confounding. In this article we provide a conceptual framework and formalization for structural nested models in continuous time. We show that the resulting estimators are consistent and asymptotically normal. Moreover, as conjectured in Robins [(1998b) *Encyclopedia of Biostatistics* **6** 4372–4389], a test for whether treatment affects the outcome of interest can be performed without specifying a model for treatment effect. We illustrate the ideas in this article with an example.

1. Introduction. Causality is a topic which nowadays receives much attention. Statisticians, epidemiologists, biostatisticians, social scientists, computer scientists [especially those in artificial intelligence, see, e.g., Pearl (2000)], econometricians and philosophers are investigating questions like "what would have happened if" and "what would happen if." This article discusses estimating the effect of a time-varying treatment. As a recurring example, this article focuses on the effect of a medical treatment which is adapted to a patient's state during the course of time.

Large observational studies have become widely used in medical research when data from randomized experiments are not available. Randomized clinical trials are often expensive, impractical, and sometimes unfeasible for ethical reasons because treatment is withheld from some patients regardless of medical considerations. Also, in some instances, exploratory investigations using nonexperimental data are used before conducting a randomized trial. In observational studies there

Received June 2004; revised March 2007.
[1]Supported in part by NIH Grants R37AI032475 and R01AI51164.
AMS 2000 subject classifications. Primary 62P10; secondary 62M99.
Key words and phrases. Causality in continuous time, counterfactuals, longitudinal data, observational studies.

is no pre-specified treatment protocol. Data are collected on patient characteristics and treatments in the course of the normal interaction between patients and doctors. Obviously, it is considerably more difficult to draw correct causal conclusions from observational data than from a randomized experiment. The main reason is the so-called confounding by indication or selection bias. For example, doctors may prescribe more medication to patients who are relatively unhealthy. Thus, association between medication dose and health outcomes may arise not only from the treatment effect but also from the way the treatment was assigned.

If this confounding by indication only takes place at the start of the treatment, one can condition on initial patient characteristics or covariates, such as blood pressure or number of white blood cells, in order to remove the effect of the confounding, and get meaningful estimates of the treatment effect. Linear regression, logistic regression or Cox regression can be used for this purpose. However, estimating treatment effects is more difficult if treatment decisions after the start of the treatment are adapted to the state of the patients in subsequent periods. Treatment might be influenced by a patient's state in the past, which was influenced by treatment decisions before; thus, simply conditioning on a patient's state in the past means disregarding information on the effect of past treatment. In such a case, even the well-known time-dependent Cox model does not answer the question of whether, or how, treatment affects the outcome of interest. The time-dependent Cox model studies the rate at which some event of interest happens (e.g., the patient dying), given past treatment- and covariate history. However, under time-dependent confounding, past covariate values may have been influenced by previous treatment. The net effect of treatment can thus not be derived from just this rate; see, for example, Robins (1998b), Keiding (1999) or Lok (2001).

Structural nested models, proposed in Robins (1989), Lok, Gill, van der Vaart and Robins (2004) and Robins (1992, 1998b) to solve practical problems in epidemiology and biostatistics, effectively overcome these difficulties and estimate the effect of time-varying treatments. The main assumption underlying these models is that all information the doctors used to make treatment decisions, and which is predictive of the patient'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. What data have to be collected to satisfy this assumption of no unmeasured confounding is for subject matter experts to decide. All of the past treatment- and covariate information which both (i) influences a doctor's treatment decisions and (ii) is relevant for a patient's prognosis with respect to the outcome of interest, has to be recorded. In Section 5 of Robins (1998b) and in Section 8.1 of Robins, Rotnitzky and Scharfstein (2000), a sensitivity analysis methodology for estimation of structural nested models is developed that does not assume no unmeasured confounders. Beyond treatment and covariates, the data requirements also include the measure of an outcome of interest; for example, survival time, time to clinical AIDS or CD4 count after the treatment period.

Lok et al. (2004) study structural nested models in discrete time. These models assume that changes in the values of the covariates and treatment decisions take place at finitely many deterministic times, which are the same for all patients and known in advance. Lok et al. (2004) also assume that covariates and treatment take values in a discrete space. They indicate why it is reasonable to expect consistency and asymptotic normality in discrete time, and they refer to Lok (2001) for the proofs. Gill and Robins (2001) generalize Lok et al. (2004) to covariates and treatment taking values in $\mathbb{R}^k$.

In this article we consider structural nested models in continuous time, proposed in Robins (1992, 1998b). Structural nested models in continuous time allow for both changes in the values of the covariates and treatment decisions to take place at arbitrary times for different patients. As noted in Robins (1998a), structural nested models in continuous time assume that a short duration of treatment has only a small effect on the distribution of the outcome of interest. The effect of the treatment on an individual patient may be large, but then the probability of such effect has to be small for any particular short duration of treatment (see page 7, bottom).

This article provides a conceptual framework and mathematical formalization of these practical methods, solving important outstanding problems and contributing to the causality discussion, especially for the time ordered and continuous time case. In particular, this article proves the conjectures in Robins (1998b) that structural nested models in continuous time lead to estimators which are both consistent and asymptotically normal. The proof simplifies considerably for structural nested models in discrete time (see our Discussion, Section 12). This article also proves that a test related to the score test can be used to investigate whether treatment affects the outcome of interest without specifying a model for the treatment effect.

2. Setting and notation. The setting to which structural nested models in continuous time apply is as follows. The outcome of interest, from now on called Y, is a continuous real variable. For example, the survival time of a patient, time to clinical AIDS, or CD4 count after the treatment period. We wish to estimate the effect of treatment on the outcome Y. There is some fixed time interval $[0, \tau]$, with τ a finite time, during which treatment and patient characteristics are observed for each patient. We suppose that after time τ treatment is stopped or switched to some kind of baseline treatment. In this article we assume that there is no censoring, and that the outcome Y is observed for every patient in the study. See, for example, Robins (1998b), Hernán et al. (2005) and Lok (2007) for ideas about dealing with censoring.

We denote the probability space by $(\Omega, \mathcal{F}, P)$. The covariate process describes the course of the patient characteristics, for example, the course of the blood pressure and the white blood cell count. We assume that a realization of this covariate process is a function from $[0, \tau]$ to $\mathbb{R}^d$, and that such a *sample path* is continuous from the right with limits from the left (cadlag). The covariates which *must*

be included are those which both (i) influence a doctor's treatment decisions *and* (ii) possibly predict a patient's prognosis with respect to the outcome of interest. If such covariates would not be observed, the assumption of no unmeasured confounding, mentioned in the introduction, will not hold.

For the moment consider one single patient. We write $Z(t)$ for the covariate- *and* treatment values at time t. We assume 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 cadlag. We write $\overline{Z}_t = (Z(s) : 0 \leq s \leq t)$ for the covariate- and treatment history until time t, and $\overline{\mathcal{Z}}_t$ for the space of cadlag functions from $[0, t]$ to $\mathbb{R}^m$ in which $\overline{Z}_t$ takes it values. Similarly, we write $\overline{Z}$ for the whole covariate- and treatment history of the patient on the interval $[0, \tau]$, and $\overline{\mathcal{Z}}$ for the space in which $\overline{Z}$ takes its values. In this article we choose the projection σ-algebra as the σ-algebra on $\overline{\mathcal{Z}}_t$ and $\overline{\mathcal{Z}}$; measurability of $Z(s)$ for each $s \leq t$ is then equivalent with measurability of the random variable $\overline{Z}_t$. For technical reasons, we include in Z a counter of the number of jump times of the measured treatment- and covariate process. We suppose that observations on different patients are independent.

3. Counterfactual outcomes. Structural nested models are models for relations between so-called counterfactuals. Consider for a moment just one patient. In reality this patient received a certain treatment and had final outcome Y. If his or her actual treatment had been stopped at time t, the patient's final outcome would possibly have been different. The outcome he or she would have had in that case we call $Y^{(t)}$. Of course, $Y^{(t)}$ is generally not observed, because the patient's actual treatment after t is usually different from no treatment; it is a counterfactual outcome. Instead of stopping treatment, one can also consider switching to some kind of baseline treatment, for example, standard treatment. Figure 1 illustrates the nature of counterfactual outcomes. We suppose that all counterfactual outcomes $Y^{(t)}$, for $t \in [0, \tau]$ and for all patients, are random variables on the probability space $(\Omega, \mathcal{F}, P)$.

4. No unmeasured confounding. To formalize the assumption of no unmeasured confounding, consider the history of a particular patient. Decisions of the doctors at time t may be based, in part, on recorded information on the state of the patient and treatment before t, that is, on $\overline{Z}_{t-} = (Z(s) : 0 \leq s < t)$, but not on other features predicting the outcome of the patient. In particular, given $\overline{Z}_{t-}$, changes of treatment at time t should be independent of $Y^{(t)}$, the outcome of the patient in case he or she would not have been treated after time t, given $\overline{Z}_{t-}$.

Note that $Y^{(t)}$ is an indication of the prognosis of the patient which does not depend on treatment decisions at or after time t, since it is the counterfactual outcome which we would have observed if treatment would have been stopped at time t. Only if treatment would have no effect, the observed outcome Y could play this role. This is why Robins' assumption of no unmeasured confounding demands the

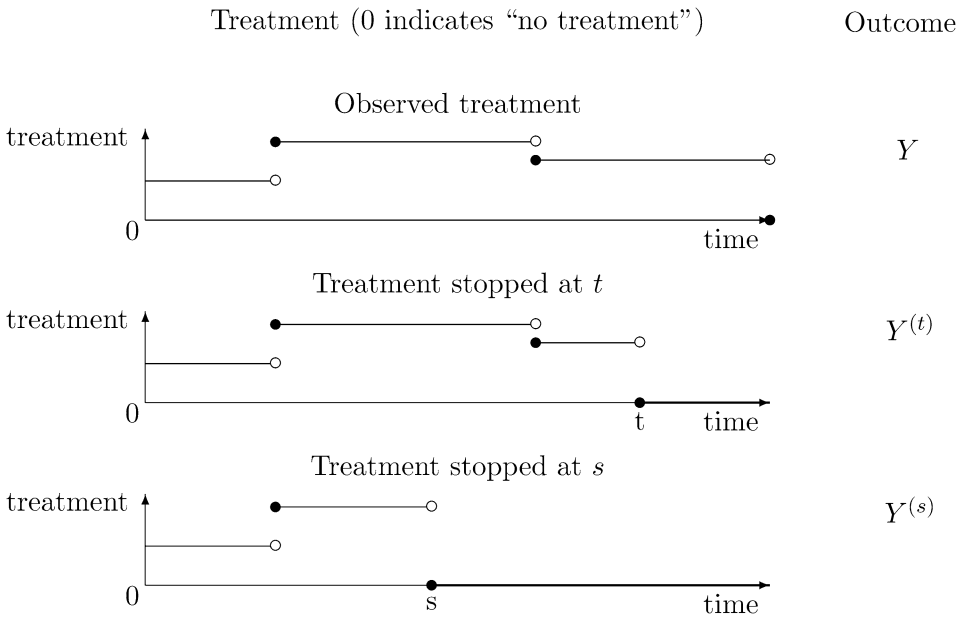

FIG. 1. *Observed and counterfactual outcomes.*

independence, given $\overline{Z}_{t-}$, of treatment decisions at time t and $Y^{(t)}$. Similar conditions, though without time-dependence, can be found in, for example, Rosenbaum and Rubin (1983).

The statement "changes of treatment at t should be independent of $Y^{(t)}$, the outcome of the patient in case he or she would not have been treated after time t, given $\overline{Z}_{t-}$" is not a formal statement: it includes conditioning of null events (since the probability that treatment changes at t may be 0 for every fixed t) on null events ($\overline{Z}_{t-}$).

To overcome this difficulty, we assume that the treatment process can be represented by or generates a (possibly multivariate) counting process N. For instance, $N(t)$ registers the number of treatment changes until time t and/or the number of times treatment reached a certain level until time t. A counting process constructed this way may serve as N in the following. More about counting processes can be found in, for example, Andersen et al. (1993). We assume that the treatment process N has an intensity process. Formally, such an intensity process $\lambda(t)$ is a predictable process such that $N(t) - \int_0^t \lambda(s)\,ds$ is a martingale. The intensity $\lambda(t)$ with respect to $\sigma(\overline{Z}_t)$ can be interpreted as the rate at which the counting process N jumps given the past treatment- and covariate history $\overline{Z}_{t-}$.

ASSUMPTION 4.1 (Bounded intensity process). N has an intensity process $\lambda(t)$ on $[0, \tau]$ with respect to the filtration $\sigma(\overline{Z}_t)$. This intensity process satisfies the following regularity conditions:

(a) λ is bounded by a constant which does not depend on ω,
(b) $\lambda(t)$ is continuous from the left.

According to this assumption,

$$M(t) = N(t) - \int_0^t \lambda(s)\,ds \tag{1}$$

is a martingale on $[0, \tau]$ with respect to the filtration $\sigma(\overline{Z}_t)$. Since most counting process martingale theory deals with filtrations $\mathcal{F}_t$ which satisfy the usual conditions ($\mathcal{F}_0$ contains all null sets and $\mathcal{F}_t = \bigcap_{s>t} \mathcal{F}_s$), we mention that, under Assumption 4.1, $M(t)$ is also a martingale with respect to $\sigma(\overline{Z}_t)^a$, the usual augmentation of $\sigma(\overline{Z}_t)$. This follows from Lemma 67.10 in Rogers and Williams (1994), since M is cadlag.

Often, N will be chosen to count the number of events of a certain type concerning the treatment process (e.g., the number of times treatment changed). At τ, the time the study ends, treatment is stopped or switched to baseline treatment, so a natural choice of N will often jump at τ with positive probability. However, jumps of N at τ are not useful for estimation, and we wish to avoid modeling jumps of N at τ. Therefore, we assume that, with probability 1, N does not jump at τ, and if a natural choice of N *does* jump at τ with positive probability, then we just adapt it, only at τ, so that it does not jump there.

We also make the following assumption.

ASSUMPTION 4.2 ($Y^{(\cdot)}$ *cadlag*). $Y^{(\cdot)}$ is a cadlag process.

Within this framework, the assumption of no unmeasured confounding could be operationalized as follows. The rate at which the counting process N jumps given past treatment- and covariate history is also the rate at which N jumps given past treatment- and covariate history and $\overline{Y}^{(t)} = (Y^{(s)} : s \le t)$. That is, the following:

ASSUMPTION 4.3 (No unmeasured confounding—formalization). The intensity process $\lambda(t)$ of N with respect to $\sigma(\overline{Z}_t)$ is also an intensity process of N with respect to $\sigma(\overline{Z}_t, \overline{Y}^{(t)})$.

This can be interpreted as conditional independence (given $\overline{Z}_{t-}$) of treatment decisions at time t and $(Y^{(s)} : s \le t)$. This assumption is stronger than just conditional independence of treatment decisions at t and $Y^{(t)}$ as assumed in Robins (1992, 1998b). However, also $Y^{(s)}$ for $s < t$ is an indication of the patient's prognosis upon which treatment decisions at time t ($> s$) should not depend. Assumption 4.3 allows us to use the usual counting processes framework. Under Assumption 4.3, $M(t) = N(t) - \int_{[0,t]} \lambda(s)\,ds$ is a martingale also with respect to $\sigma(\overline{Z}_t, \overline{Y}^{(t)})$ and its usual augmentation $\sigma(\overline{Z}_t, \overline{Y}^{(t)})^a$.

This formalization of the assumption of no unmeasured confounding in terms of compensators with respect to the filtration $\sigma(\overline{Z}_t, \overline{Y}^{(t)})$ is a novel feature of this paper relative to the previous literature on structural nested models. Robins et al. (1992), Robins (1998b) and Keiding (1999) use a Cox model for initiation and/or changes in treatment. However, none of them formalized the assumption of no unmeasured confounders in terms of compensators with respect to $\sigma(\overline{Z}_t, \overline{Y}^{(t)})$. As a consequence, they could not use the extensive theory on counting process martingales to show the asymptotics of their estimators, which then remained without proof.

5. The 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

$$D(y, t; \overline{Z}_t) = \frac{\partial}{\partial h}\bigg|_{h=0} \big(F^{-1}_{Y^{(t+h)}|\overline{Z}_t} \circ F_{Y^{(t)}|\overline{Z}_t}\big)(y), \tag{2}$$

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, $\overline{Z}_t$. Notice that for differentiability of $F^{-1}_{Y^{(t+h)}|\overline{Z}_t}$ with respect to h we need that a short duration of treatment has only a small effect on the distribution of the outcome of interest (see page 3, second paragraph), since $\lim_{h \downarrow 0} F_{Y^{(t+h)}|\overline{Z}_t}$ must be equal to $F_{Y^{(t)}|\overline{Z}_t}$.

EXAMPLE 5.1 (*Survival of AIDS patients*). Robins, Blevins, Ritter and Wulfsohn (1992) describe an AIDS clinical trial to study the effect of AZT treatment on survival in HIV-infected subjects. Embedded within this trial was an essentially uncontrolled observational study of the effect of prophylaxis therapy for PCP on survival. PCP, pneumocystis carinii pneumonia, is an opportunistic infection that afflicts AIDS patients. The aim of Robins et al. (1992) was to study the effect of this prophylaxis therapy on survival. Thus, the outcome of interest Y is the survival time, and the treatment under study is prophylaxis for PCP. Once treatment with prophylaxis for PCP started, it was never stopped.

Suppose that

$$D(y, t; \overline{Z}_t) = (1 - e^{\psi}) 1_{\{\text{treated at } t\}}. \tag{3}$$

Then (see Section 6 for details), for $t < Y$, withholding treatment from t onward leads to (with $\sim$ meaning "is distributed as")

$$\begin{aligned} Y^{(t)} - t &\sim \int_t^Y e^{\psi 1_{\{\text{treated at } s\}}}\, ds \\ &= e^{\psi} \cdot DUR(t, Y) + 1 \cdot \big(Y - t - DUR(t, Y)\big) \qquad \text{given } \overline{Z}_t, \end{aligned} \tag{4}$$

with $DUR(t,u)$ the duration of treatment in the interval (t,u). Thus, treated residual survival time (t until Y) is multiplied by e^{ψ} by withholding treatment; compare this with accelerated failure time models, see, for example, Cox and Oakes (1984). This multiplication factor e^{ψ} should be interpreted in a distributional way. One of the models studied in Robins et al. (1992) assumes that (4) is true even with $\sim$ replaced by $=$ (though only for $t=0$). Notice that supposing (4) to be true with $\sim$ replaced by $=$ would be much stronger.

EXAMPLE 5.2 (*Survival of AIDS patients*). Consider the situation from Example 5.1 again. In another model mentioned in Robins, Blevins, Ritter and Wulfsohn (1992) the factor with which treated residual survival time is multiplied when treatment is withheld can depend on the AZT treatment the patient received and whether or not the patient had a history of PCP prior to start of PCP prophylaxis. Since this was a clinical trial for AZT treatment, the AZT treatment was described by a single variable I_{AZT} indicating the treatment arm the patient was randomized to (I_{AZT} is 0 or 1). Whether or not the patient had a PCP history prior to start of prophylaxis is described by an indicator variable $P(t)$. $P(t)$ equals 1 if the patient had PCP before or at t *and* before prophylaxis treatment started; otherwise $P(t)$ equals 0. If

$$D_{\psi_1,\psi_2,\psi_3}(y,t;\overline{Z}_t) = \bigl(1-e^{\psi_1+\psi_2 P(t)+\psi_3 I_{\mathrm{AZT}}}\bigr)1_{\{\text{treated at } t\}},$$

then (see Section 6 for details) withholding prophylaxis treatment from t onward leads to

$$Y^{(t)} - t \sim \int_t^Y e^{1_{\{\text{treated at } s\}}(\psi_1+\psi_2 P(s)+\psi_3 I_{\mathrm{AZT}})}\,ds \qquad \text{given } \overline{Z}_t, \tag{5}$$

for $t < Y$.

EXAMPLE 5.3 (*Incorporating a-priori biological knowledge*). Following Robins (1998b), again consider survival as the outcome of interest. Suppose that it is known that treatment received at time t only affects survival for patients destined to 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 subject is already infected and, if death occurs, it always does within five weeks from the time of initial unrecorded subclinical infection. In that case, as remarked in Robins (1998b), the natural restriction on D is that

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

More biostatistical examples of models for D can be found in, for example, Mark and Robins (1993), Witteman et al. (1998), Robins (1998b) and Keiding et al. (1999).

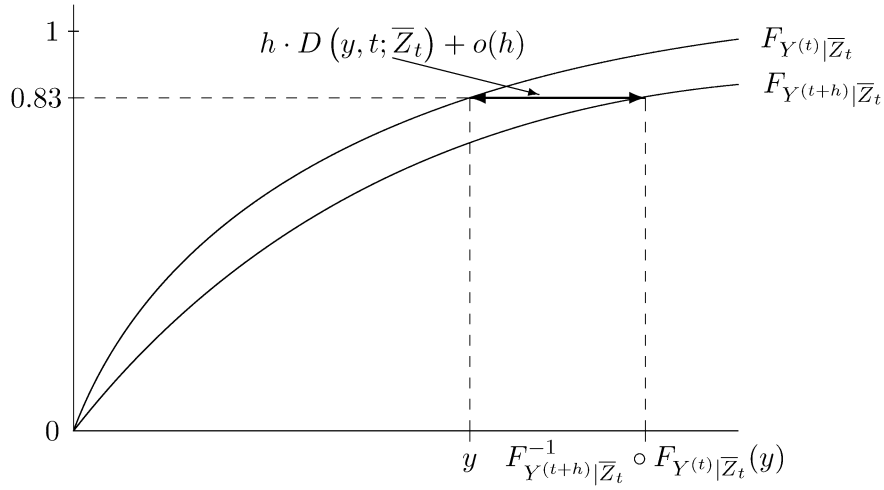

FIG. 2. *Illustration of the infinitesimal shift-function D.*

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

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

In Figure 2 this is sketched. y in the picture is the 0.83th quantile of the distribution of $Y^{(t)}$ given $\overline{Z}_t$. For $h > 0$, the 0.83th quantile of the distribution of $Y^{(t+h)}$ given $\overline{Z}_t$ is $y + h \cdot D(y, t; \overline{Z}_t) + o(h)$. Thus, to shift from quantiles of the distribution of $Y^{(t)}$ to the distribution of $Y^{(t+h)}$ given $\overline{Z}_t$ ($h > 0$) is approximately the same as to just add $h \cdot D(y, t; \overline{Z}_t)$ to those quantiles. For example, if $F_{Y^{(t+h)}|\overline{Z}_t}$ lies to the right of $F_{Y^{(t)}|\overline{Z}_t}$ for $h > 0$, then treatment between t and $t+h$ increases the outcome (in distribution), and $D(\cdot, t; \overline{Z}_t)$ is greater than 0.

Consider again this interpretation of D as the infinitesimal effect of treatment given in $[t, t+h)$. If the outcome of interest is survival, then $D(y, t; \overline{Z}_t)$ should be zero if $\overline{Z}_t$ indicates the patient is dead at time t. Indeed, in that case $F_{Y^{(t+h)}|\overline{Z}_t}$ and $F_{Y^{(t)}|\overline{Z}_t}$ should be almost surely the same for every $h \geq 0$, since withholding treatment after death does not change the survival time. Thus, $F^{-1}_{Y^{(t+h)}|\overline{Z}_t} \circ F_{Y^{(t)}|\overline{Z}_t}(y)$ is constant in h for $h \geq 0$ and, therefore, $D(y, t; \overline{Z}_t) = 0$. However, this reasoning is not precise because of the complication of null sets. We will therefore just formally define $D(y, t; \overline{Z}_t)$ to be zero if the outcome of interest is survival and $\overline{Z}_t$ indicates the patient is dead at time t.

It can be shown that $D \equiv 0$ if treatment does not affect the outcome of interest, as was conjectured in Robins (1998b). To be more precise, Lok (2001) 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

patient characteristics are selected at that time ($\overline{Z}_t$), stopping 'treatment as given' at some fixed time after t would not change the distribution of the outcome in patients with these patient characteristics."

In the rest of this article D_ψ will always indicate a correctly specified parametric model for D, with $D \equiv 0$ if $\psi = 0$.

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

$$X'(t) = D(X(t), t; \overline{Z}_t) \tag{6}$$

with final condition $X(\tau) = Y$, the observed outcome (see Figure 3). Then $X(t)$ mimics $Y^{(t)}$ in the sense that it has the same distribution as $Y^{(t)}$ given $\overline{Z}_t$. This rather surprising result was conjectured in Robins (1998b) and proved in Lok (2001, 2004). To prove this result, we need the following consistency assumption.

ASSUMPTION 6.1 (Consistency). $Y^{(\tau)}$ has the same distribution as Y given $\overline{Z}_\tau$.

Notice that since by assumption no treatment was given after time τ and since treatment is right-continuous, there is no difference in treatment between $Y^{(\tau)}$

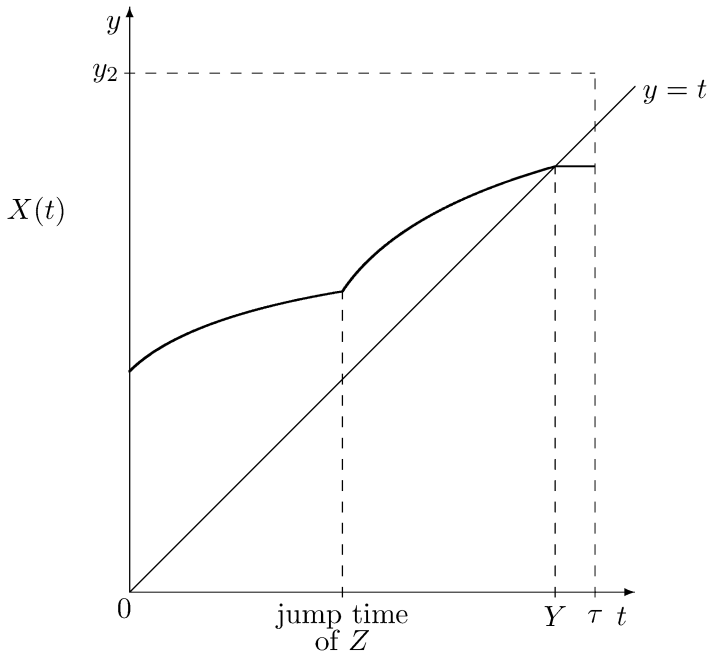

FIG. 3. *An example of a solution $X(t)$ to the differential equation $X'(t) = D(X(t), t; \overline{Z}_t)$ with final condition $X(\tau) = Y$ in case the outcome is survival time.*

and Y. We suppose that this assumption holds, and we also suppose that a short duration of treatment has only a small effect on the distribution of the outcome of interest ($\lim_{h\downarrow 0} F_{Y^{(t+h)}|\overline{Z}_t}(y) \to F_{Y^{(t)}|\overline{Z}_t}(y)$). Under these assumptions and regularity conditions only, Lok (2001, 2004) proved that indeed equation (6) has a unique solution for every $\omega \in \Omega$, and that this solution $X(t)$ mimics $Y^{(t)}$ in the sense that $X(t)$ has the same distribution as $Y^{(t)}$ given $\overline{Z}_t$ (see Appendix B). Throughout this article we will assume that this result holds true.

EXAMPLE 6.2. Survival of AIDS patients (continuation of Example 5.1). Suppose that

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

Then

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

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

Suppose now that one has a correctly specified parametric model for the infinitesimal shift-function D, D_ψ. Then one can calculate "$X_\psi(t)$," the solution to

$$X'_\psi(t) = D_\psi(X_\psi(t), t; \overline{Z}_t) \tag{7}$$

with final condition $X_\psi(\tau) = Y$. For the true ψ, $X_\psi(t)$ has the same distribution as $Y^{(t)}$, the outcome with treatment stopped at t, even given all patient-information at time t, $\overline{Z}_t$. So instead of the unobservable $Y^{(t)}$'s, we have the observable $X_\psi(t)$'s which for the true ψ mimic the $Y^{(t)}$'s. Although we do not know the true ψ, this result turns out to be very useful, both for estimating ψ (Sections 8 and 10) and for testing [Section 11; notice that when testing whether treatment affects the outcome (i.e., whether $D \equiv 0$), X can simply be calculated from the data ($X \equiv Y$) under the null hypothesis of no treatment effect].

7. Local rank preservation. Previous applications of structural nested models [see, e.g., Robins et al. (1992), Mark and Robins (1993), Witteman et al. (1998) and Keiding et al. (1999)] have assumed the so-called local rank preservation condition. Local rank preservation states that $Y^{(t)}$ is a local solution to (6). However, if $Y^{(t)}$ is locally a solution to (6), it is usually also globally a solution to (6); see, for example, Theorem A.1 in the Appendix. Hence, if one knew the parameter ψ, every $Y^{(t)}$ would be a deterministic function of the observed data. Deterministic dependence of counterfactuals on the observed data is a very strong condition, which, though untestable, is generally considered implausible. The previous literature [see, e.g., Robins (1998b), Robins et al. (1992), Mark and Robins (1993) and Keiding et al. (1999)] acknowledged this problem, and conjectured that the assumption of local rank preservation could be relaxed in continuous time [since

it is known that the assumption of local rank preservation can be relaxed for structural nested models in discrete time; this was pointed out by Robins and Wasserman (1997), and Lok et al. (2004) provided a proof]. See Robins (1998b) for a more elaborate discussion. The following example describes a setting where the assumption of local rank preservation is implausible.

EXAMPLE 7.1 (*Survival of AIDS patients and local rank preservation*). In the situation of Example 5.1, consider the following thought experiment. Suppose that two patients had the same covariate history until time t, and both received the same constant treatment to prevent PCP until time t (equal $\overline{Z}_t$). Suppose, furthermore, that both patients received no treatment after time t, that they did not have PCP before time t and both died at the same time $u > t$ (for both, $Y^{(t)} = u$). Possibly, the first patient would have had PCP at some time $s < t$ and would have died from it before u in case he or she would not have been treated. Possibly, the other patient would not have had PCP in case he or she would not have been treated, and would have died at the same time $u > t$ as without treatment. Thus, it is easy to imagine that these patients would have had different outcomes under no treatment (different $Y^{(0)}$). However, the assumption of local rank preservation excludes this possibility.

Local rank preservation is a very strong condition, for which structural nested models have previously been attacked. In fact, this article shows that the assumption of local rank preservation is not needed for structural nested models. However, proofs would be much easier under rank preservation; for details, see the remarks before the proofs of Theorems 8.5 and 9.2. See also Robins (1998b) for a more informal reasoning.

8. Estimation of treatment effect. To estimate the infinitesimal shift-function D, Robins (1998b) proposes to use a (semi-)parametric model to predict future treatment (N in our case) on the basis of past treatment- and covariate history $\overline{Z}_{t-}$. This may seem odd, since prediction of treatment is not what we are interested in. However, we will show that such a model to predict treatment changes can indeed be a tool to get unbiased estimating equations for the parameter ψ in the model for D. Moreover, often doctors may have a better understanding, at least qualitatively, about how decisions about treatment were made than about the effect of the treatment. In what follows we will assume that λ_θ is a correctly specified parametric model for the intensity λ of N.

Recall from Section 4 that, under no unmeasured confounding (Assumption 4.3), $Y^{(t)}$ does not contain information about treatment changes given past treatment- and covariate history $\overline{Z}_{t-}$. Since $X(t)$ has the same distribution as $Y^{(t)}$ given $\overline{Z}_t$, one could expect that also $X(t)$ does not contain information about treatment changes given $\overline{Z}_{t-}$. Unfortunately, this reasoning is not precise: we have to somehow deal with null sets since the probability that treatment changes at t given

past covariate- and treatment history is often equal to 0 for each t. In Section 9 we will show how this can be dealt with.

In the current section we present a class of unbiased estimating equations for θ and ψ. These will be used for the proof in the next section, but they are also of interest in their own. In Section 9 we will see that these estimating equations are in fact martingales, for the true parameters θ_0 and ψ_0.

Recall from Section 4 that, under no unmeasured confounding, we have the martingale $M(t) = N(t) - \int_{[0,t]} \lambda(s)\,ds$ with respect to the filtration $\sigma(\overline{Z}_t, \overline{Y}^{(t)})$ and its usual augmentation. From this martingale we can construct a whole family of martingales. If $h_t(Y^{(t-)}, \overline{Z}_{t-})$ is a $\sigma(\overline{Z}_t, \overline{Y}^{(t)})^a$-predictable process, then, under regularity conditions,

$$\int_0^t h\,dM = \int_0^t h_s\big(Y^{(s-)}, \overline{Z}_{s-}\big)\,dM(s)$$

is a martingale with respect to $\sigma(\overline{Z}_t, \overline{Y}^{(t)})^a$. For a more formal statement, we first make sure that $h_t(Y^{(t-)}, \overline{Z}_{t-})$ is predictable. We put the following restriction on the functions h_t we consider here:

RESTRICTION 8.1. When in this section we consider functions h_t from $\mathbb{R} \times \overline{\mathcal{Z}}_{t-}$, we assume that they are measurable and satisfy the following:

(a) h_t is bounded by a constant which does not depend on t and $\overline{Z}$,

(b) for all $t_0 \in [0, \tau]$, $y_0 \in \mathbb{R}$ and $\omega \in \Omega$, $h_t(y, \overline{Z}_{t-}(\omega)) \to h_{t_0}(y_0, \overline{Z}_{t_0-}(\omega))$ when $y \to y_0$ and $t \uparrow t_0$.

For such h_t, $h_t(Y^{(t-)}, \overline{Z}_{t-})$ is a $\sigma(\overline{Z}_t, \overline{Y}^{(t)})^a$-predictable process:

LEMMA 8.2. *Suppose that $Y^{(\cdot)}$ is cadlag (Assumption 4.2). Then $h_t(Y^{(t-)}, \overline{Z}_{t-})$ is a $\sigma(\overline{Z}_t, \overline{Y}^{(t)})^a$-predictable process for any h_t satisfying Restriction 8.1.*

PROOF. $h_t(Y^{(t-)}, \overline{Z}_{t-})$ is adapted. It is also left-continuous: because $Y^{(\cdot)}$ is cadlag, $\lim_{t \uparrow t_0} Y^{(t-)} = Y^{(t_0-)}$ exists. □

Thus, we come to the following lemma:

LEMMA 8.3. *Under Assumptions 4.1 (bounded intensity process), 4.2 ($Y^{(\cdot)}$ cadlag) and 4.3 (no unmeasured confounding),*

$$\int_0^t h_s\big(Y^{(s-)}, \overline{Z}_{s-}\big)\big(dN(s) - \lambda(s)\,ds\big)$$

is a martingale on $[0, \tau]$ with respect to $\sigma(\overline{Z}_t, \overline{Y}^{(t)})^a$ for all h_t satisfying Restriction 8.1.

PROOF. $M(t) = N(t) - \int_0^t \lambda(s)\,ds$ is a martingale on $[0, \tau]$ with respect to $\sigma(\overline{Z}_t, \overline{Y}^{(t)})^a$, because of Assumption 4.3. It is of integrable variation [$E\int_0^t |dM(s)| \le E\int_0^t dN(s) + \lambda(s)\,ds = 2E\int_0^t \lambda(s)\,ds$, and λ is bounded (Assumption 4.1)]. Because of Lemma 8.2, $h(t) = h_t(Y^{(t-)}, \overline{Z}_{t-})$ is a $\sigma(\overline{Z}_t, \overline{Y}^{(t)})^a$-predictable process. It is also bounded [Restriction 8.1(a)]. Hence, $\int_0^t h(s)\,dM(s) = \int_0^t h_s(Y^{(s-)}, \overline{Z}_{s-})(dN(s) - \lambda(s)\,ds)$ is an integral of a bounded predictable process with respect to a martingale of integrable variation, and, therefore, a $\sigma(\overline{Z}_t, \overline{Y}^{(t)})^a$-martingale. □

To construct unbiased estimating equations for (θ_0, ψ_0), we need to assume that the probability that $N(\cdot)$ and $Y^{(\cdot)}$ jump at the same time is zero. This assumption is a formalization of the assumption of no instantaneous treatment effect as proposed in Robins (1998b), which can be seen as follows. Given $\overline{Z}_{t-}$ and $\overline{Y}^{(t-)}$, N jumps at t with rate $\lambda(t)$ (Assumption 4.3, no unmeasured confounding). $Y^{(\cdot)}$ is a cadlag process (Assumption 4.2), which thus for every $\omega \in \Omega$ jumps at most countably many times on the finite time interval $[0, \tau]$. Therefore, if $Y^{(\cdot)}$ and N would jump at the same time with positive probability, this would imply a dependence of these jumps; the obvious interpretation of this dependence would be that a change of treatment instantaneously affects the outcome of interest.

ASSUMPTION 8.4 (*No instantaneous treatment effect*). The probability that there exists a t such that $N(\cdot)$ and $Y^{(\cdot)}$ both jump at time t is 0.

Notice that this excludes estimation of the effect of point exposures. For example, if treatment is surgery or another point exposure given at some time t, the outcome under "treatment stopped at time t" will typically jump at time t if treatment affects the outcome of interest, at the same time as the treatment itself. However, this assumption does not exclude the possibility that the outcome differs depending on whether a patient is treated or not at a certain point in time. For example, $Y^{(t+)}$ and $Y^{(t-)}$ may be different when a virus is contacted at time t. The model in this article can accommodate differences between $Y^{(t+)}$ and $Y^{(t-)}$, as long as the probability that the observed treatment changes is 0 at that precise time. Or, in more generality, as long as the probability that N jumps at the same time is 0. This was previously noticed in Robins (1998a), Section 8. The estimating procedures in this article do not deal with instantaneous treatment effects.

Suppose that the above conditions hold and that $(X(t), \overline{Z}_t) \sim (Y^{(t)}, \overline{Z}_t)$ for $t \in [0, \tau]$ (see Section 6). Then if D_ψ and λ_θ are correctly specified (parametric) models for D and λ, respectively, each choice of h_t satisfying Restriction 8.1 leads to an unbiased estimating equation for both the parameter of interest ψ and the (nuisance) parameter θ:

THEOREM 8.5. *Suppose that Assumptions* 4.1 (*bounded intensity process*), 4.2 ($Y^{(\cdot)}$ *cadlag*), 4.3 (*no unmeasured confounding*) *and* 8.4 (*no instantaneous treatment effect*) *are satisfied. Suppose also that, for every* $t \in [0, \tau]$, $X(t)$ *has the same distribution as* $Y^{(t)}$ *given* $\overline{Z}_t$. *Then*

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

for each h_t *satisfying Restriction* 8.1. *Thus, if* D_ψ *and* λ_θ *are correctly specified parametric models for* D *and* λ, *respectively,*

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

with P_n *the empirical measure* $P_n X = 1/n \sum_{i=1}^n X_i$, *is an unbiased estimating equation for* (θ_0, ψ_0), *for each* h_t *satisfying Restriction* 8.1. h_t *here is allowed to depend on* ψ *and* θ, *as long as it satisfies Restriction* 8.1 *for* (θ_0, ψ_0).

As before, $X_\psi(t)$ here is the continuous solution of (7), $X'_\psi(t) = D_\psi(X_\psi(t), t; \overline{Z}_t)$ with boundary condition $X_\psi(\tau) = Y$. Moreover, as before, we assume that for all D_ψ we have existence and uniqueness of such solutions; Theorem A.1 in the appendix provides sufficient conditions for that.

Under local rank preservation (see Section 7), $X(t) = Y^{(t)}$ for each t. In that case Theorem 8.5 follows immediately from Lemma 8.3. However, as argued in Section 7, local rank preservation is generally considered implausible.

PROOF OF THEOREM 8.5. We have to show that

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

has expectation zero for all h_t satisfying Restriction 8.1. To do that, we prove that it has the same expectation as

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

which has expectation zero because of Lemma 8.3. We will first show that the terms with dN have the same expectation, that is,

$$E\left(\sum_{t \le \tau, \Delta N(t) = 1} h_t(X(t), \overline{Z}_{t-})\right) = E\left(\sum_{t \le \tau, \Delta N(t) = 1} h_t\big(Y^{(t-)}, \overline{Z}_{t-}\big)\right). \tag{8}$$

After that we show that the terms with $\lambda(t)\, dt$ have the same expectation, that is,

$$E\left(\int_0^\tau h_t(X(t), \overline{Z}_{t-}) \lambda(t)\, dt\right) = E\left(\int_0^\tau h_t\big(Y^{(t-)}, \overline{Z}_{t-}\big) \lambda(t)\, dt\right). \tag{9}$$

As we will see below, (8) and (9) have to be proved separately, since we do not have or expect that $(X(s), \overline{Z}_t) \sim (Y^{(s)}, \overline{Z}_t)$ for $s < t$; we only have this for $s \geq t$. Therefore, the approximations below have to be chosen carefully.

At first we prove (8), by approximating these sums and showing that the approximations have the same expectation. Next we show that the approximations converge and that (8) follows with Lebesgue's dominated convergence theorem.

Define $T_1 = \inf\{t : N(t) = 1\}$, $T_2 = \inf\{t : N(t) = 2\}$, etc., the jump times of the counting process N in the interval $[0, \tau]$. They are measurable [e.g., because of Rogers and Williams (1994), Lemma 74.4]. Note that the number of jumps in $[0, \tau]$ is almost surely finite because N is integrable (it has a bounded intensity process). In the following read $h_{T_j}(Y^{(T_j-)}, \overline{Z}_{T_j-}) = 0$ if there is no jth jump of N in the interval $[0, \tau]$.

Next split up the interval $[0, \tau]$ in intervals of equal length: for $K \in \mathbb{N}$ fixed, put $\tau_k = k\tau/K$, $k = 0, \ldots, K$. Fix K for the moment. The right-hand side of equation (8) is harder to approximate than the left-hand side, both because $Y^{(t)}$ does not need to be continuous in t while $X(t)$ does and because knowing $Y^{(t)}$ and $\overline{Z}_t$ does not imply knowing $Y^{(s)}$ for $s < t$ and we do not have or expect $(X(s), \overline{Z}_t) \sim (Y^{(s)}, \overline{Z}_t)$ for $s < t$. The approximations we choose are

$$\begin{aligned}\sum_{\Delta N(t)=1, t\leq\tau} h_t\bigl(Y^{(t-)}, \overline{Z}_{t-}\bigr) &= \sum_{j=1}^{\infty} h_{T_j}\bigl(Y^{(T_j-)}, \overline{Z}_{T_j-}\bigr) \\ &\approx \sum_{j=1}^{\infty} \sum_{k=0}^{K-1} 1_{(\tau_k, \tau_{k+1}]}(T_j) h_{\tau_k}\bigl(Y^{(\tau_{k+1})}, \overline{Z}_{\tau_k-}\bigr)\end{aligned} \tag{10}$$

and

$$\begin{aligned}\sum_{\Delta N(t)=1, t\leq\tau} h_t(X(t), \overline{Z}_{t-}) &= \sum_{j=1}^{\infty} h_{T_j}(X(T_j), \overline{Z}_{T_j-}) \\ &\approx \sum_{j=1}^{\infty} \sum_{k=0}^{K-1} 1_{(\tau_k, \tau_{k+1}]}(T_j) h_{\tau_k}\bigl(X(\tau_{k+1}), \overline{Z}_{\tau_k-}\bigr).\end{aligned} \tag{11}$$

To show that these approximations have the same expectation, we use that $(X(\tau_{k+1}), \overline{Z}_{\tau_{k+1}}) \sim (Y^{(\tau_{k+1})}, \overline{Z}_{\tau_{k+1}})$. Therefore, also

$$1_{(\tau_k, \tau_{k+1}]}(T_j) h_{\tau_k}(X(\tau_{k+1}), \overline{Z}_{\tau_k-}) \sim 1_{(\tau_k, \tau_{k+1}]}(T_j) h_{\tau_k}\bigl(Y^{(\tau_{k+1})}, \overline{Z}_{\tau_k-}\bigr)$$

[notice that $1_{(\tau_k, \tau_{k+1}]}(T_j)$ is a function of $\overline{Z}_{\tau_{k+1}}$]. Hence, the expectation of each of the terms on the right-hand side of (10) is equal to the expectation of the corresponding term on the right-hand side of (11). Since h_t is bounded [Restriction 8.1(a)] and the expected number of jump times T_j is finite (N is integrable), this implies that the expectation of the right hand-side of equation (10) is equal to the expectation of the right-hand side of equation (11).

Equation (8) follows if the expectation of the approximations in (10) and (11) converges to the right-hand side and left-hand side of equation (8), respectively. This convergence is harder to show for (10) than for (11), since $Y^{(\cdot)}$ may jump [while $X(\cdot)$ does not, by construction]. Fix j for a moment. Define τ_k^j and τ_{k+1}^j as the grid points such that $T_j \in (\tau_k^j, \tau_{k+1}^j]$. As $K \to \infty$, $\tau_{k+1}^j \downarrow T_j$, so that since $Y^{(\cdot)}$ is cadlag, $Y^{(\tau_{k+1}^j)} \to Y^{(T_j)}$. Moreover, as $K \to \infty$, $\tau_k^j \uparrow T_j$, so that because of Restriction 8.1 on h, $h_{\tau_k^j}(Y^{(\tau_{k+1}^j)}, \overline{Z}_{\tau_k^j-}) \to h_{T_j}(Y^{(T_j)}, \overline{Z}_{T_j-})$. Combining this for all j leads to

$$\sum_{j=1}^{\infty}\sum_{k=0}^{K-1} 1_{(\tau_k,\tau_{k+1}]}(T_j) h_{\tau_k}\bigl(Y^{(\tau_{k+1})}, \overline{Z}_{\tau_k-}\bigr) \to \sum_{j=1}^{\infty} h_{T_j}\bigl(Y^{(T_j)}, \overline{Z}_{T_j-}\bigr)$$

as $K \to \infty$ for every ω for which the number of jumps of N is finite, so for almost every $\omega \in \Omega$. h_t is bounded [Restriction 8.1(a)] and the left-hand side is bounded by the number of jumps of N times this bound. The expectation of that is finite because N is integrable. Thus, Lebesgue's dominated convergence theorem can be applied, and

$$E\left(\sum_{j=1}^{\infty}\sum_{k=0}^{K-1} 1_{(\tau_k,\tau_{k+1}]}(T_j) h_{\tau_k}\bigl(Y^{(\tau_{k+1})}, \overline{Z}_{\tau_k-}\bigr)\right) \to E\left(\sum_{j=1}^{\infty} h_{T_j}\bigl(Y^{(T_j)}, \overline{Z}_{T_j-}\bigr)\right) \tag{12}$$

as $K \to \infty$. Because with probability one $Y^{(\cdot)}$ and N do not jump at the same time (Assumption 8.4 of no instantaneous treatment effect),

$$\sum_{j=1}^{\infty} h_{T_j}\bigl(Y^{(T_j)}, \overline{Z}_{T_j-}\bigr) = \sum_{j=1}^{\infty} h_{T_j}\bigl(Y^{(T_j-)}, \overline{Z}_{T_j-}\bigr) \qquad \text{a.s.},$$

so that we can replace $Y^{(T_j)}$ by $Y^{(T_j-)}$ on the right-hand side of (12). Therefore, indeed, the expectation of the approximation in (10) converges to the expectation of the left-hand side of (10). The same reasoning shows this for (11). Here less caution is necessary since $X(t)$ is continuous in t. That concludes the proof of equation (8).

Next we prove (9), also by approximation. Here, too, we show that the approximations have the same expectation and that (9) follows with Lebesgue's dominated convergence theorem.

Divide the interval $[0, \tau]$ as above. The approximations we choose here are

$$\int_0^{\tau} h_t\bigl(Y^{(t-)}, \overline{Z}_{t-}\bigr)\lambda(t)\,dt \approx \sum_{k=0}^{K-1} h_{\tau_k}\bigl(Y^{(\tau_k)}, \overline{Z}_{\tau_k-}\bigr)\lambda(\tau_k)(\tau_{k+1} - \tau_k) \tag{13}$$

and

$$\int_0^{\tau} h_t(X(t), \overline{Z}_{t-})\lambda(t)\,dt \approx \sum_{k=0}^{K-1} h_{\tau_k}(X(\tau_k), \overline{Z}_{\tau_k-})\lambda(\tau_k)(\tau_{k+1} - \tau_k). \tag{14}$$

Because $(X(\tau_k), \overline{Z}_{\tau_k}) \sim (Y^{(\tau_k)}, \overline{Z}_{\tau_k})$ and $\lambda(\tau_k)$ are a measurable function of $\overline{Z}_{\tau_k}$ (Assumption 4.1, bounded intensity process), the expectation of each of the terms in (13) is equal to the expectation of the corresponding term in (14). Thus, the expectations of these approximations are equal.

Equation (9) follows if the expectation of the approximations in (13) and (14) converge to the right-hand side and left-hand side of equation (9), respectively. This convergence is also harder to show for (13) than for (14) because of possible discontinuities of $Y^{(\cdot)}$. First notice that as $K \to \infty$, for t fixed,

$$\sum_{k=0}^{K-1} 1_{(\tau_k, \tau_{k+1}]}(t) h_{\tau_k}\bigl(Y^{(\tau_k)}, \overline{Z}_{\tau_k -}\bigr)\lambda(\tau_k) \to h_t\bigl(Y^{(t-)}, \overline{Z}_{t-}\bigr)\lambda(t)$$

for every $\omega \in \Omega$ fixed and for every $t < \tau$: $Y^{(\cdot)}$ has limits from the left (Assumption 4.2), so that as $\tau_k \uparrow t$, Restriction 8.1(b) on h can be used, and λ is continuous from the left [Assumption 4.1(b)]. Taking integrals and applying Lebesgue's dominated convergence theorem [h_t and λ are bounded because of Restriction 8.1(a) and Assumption 4.1(a), resp.] leads to

$$\sum_{k=0}^{K-1} h_{\tau_k}\bigl(Y^{(\tau_k)}, \overline{Z}_{\tau_k -}\bigr)\lambda(\tau_k)(\tau_{k+1} - \tau_k) \to \int_0^{\tau} h_t\bigl(Y^{(t-)}, \overline{Z}_{t-}\bigr)\lambda(t)\,dt$$

for every $\omega \in \Omega$. As both h and λ are bounded, Lebesgue's dominated convergence theorem guarantees that indeed the expectation of the approximation in (13) converges to the expectation of the left-hand side of (13). The same reasoning shows this for (14), which concludes the proof of equation (9) and Theorem 8.5. □

Lok (2001) shows that if the rest of the conditions in this section are satisfied, Assumption 8.4 (treatment does not instantaneously affect the outcome of interest) is a necessary condition for Theorem 8.5.

EXAMPLE 8.6 (Survival of AIDS patients and the Weibull proportional hazards model). Consider the setting of Examples 5.1 and 5.2 and define $N(t) = 1$ if prophylaxis treatment started at or before time t and 0 otherwise. Suppose that initiation of prophylaxis treatment can be correctly modeled with the time-dependent Weibull proportional hazards model

$$\lambda_{\xi,\gamma,\theta}(t) = 1_{\{\text{at risk at } t\}}\,\xi\gamma t^{\gamma-1} e^{\theta_1 I_{\mathrm{AZT}} + \theta_2 I_{\mathrm{PCP}}(t)},$$

where $I_{\mathrm{PCP}}(t)$ equals 1 if the patient had PCP before time t and 0 otherwise, and ξ and γ are greater than zero [for more about the Weibull proportional hazards model and its applications see, e.g., Collett (1994)]. If the patient died before t

or prophylaxis treatment already started before, the patient is not "at risk" for initiation of treatment and, thus, λ equals 0. Then the (partial) score equations for estimation of (ξ, γ, θ) are

$$P_n \int_0^\tau \begin{pmatrix} \frac{1}{\xi} & \frac{1}{\gamma} + \log t & I_{\mathrm{AZT}} & I_{\mathrm{PCP}}(t) \end{pmatrix}^\top \big(dN(t) - \lambda_{\xi,\gamma,\theta}(t)\,dt\big) = 0.$$

Such estimating equations can also be written down for the model including αX_ψ,

$$\lambda_{\xi,\gamma,\theta,\alpha,\psi}(t) = 1_{\{\text{at risk at } t\}}\, \xi\gamma t^{\gamma-1} e^{\theta_1 I_{\mathrm{AZT}} + \theta_2 I_{\mathrm{PCP}}(t) + \alpha X_\psi(t)}.$$

Robins (1998b) proposes to estimate the parameters in a model like this by choosing those parameters $(\xi, \gamma, \theta, \psi)$ which maximize the likelihood when X_ψ is considered fixed and known, and for which $\hat{\alpha}(\psi) = 0$: for the true ψ, $X_\psi(t) = X(t) \sim Y^{(t)}$ does not contribute to the model for treatment changes (under no unmeasured confounding). To make the connection with the estimators in the current article, notice that this leads to the same estimators as the ones that solve the estimating equations arising from the likelihood when X_ψ is considered fixed and known, with α put to zero. More precise, since we know that the true α is equal to 0, we put α equal to 0 and get the estimating equations

$$P_n \int_0^\tau \begin{pmatrix} \frac{1}{\xi} & \frac{1}{\gamma} + \log t & I_{\mathrm{AZT}} & I_{\mathrm{PCP}}(t) & X_\psi(t) \end{pmatrix}^\top \big(dN(t) - \lambda_{\xi,\gamma,\theta}(t)\,dt\big) = 0$$

for the parameter ψ (and thus also for D) and the (nuisance) parameters (ξ, γ, θ). These estimating equations are of the form of Theorem 8.5,

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

but the function h_t here is not bounded and λ need not be bounded (if $\gamma < 1$), so unbiasedness does not follow immediately from Theorem 8.5. However, we could restrict the interval $[0, \tau]$ to $[\varepsilon, \tau]$ for $\varepsilon > 0$ (to assure that λ is bounded) and $\log t$ can be approximated by the bounded functions $\log t \vee C$ $(C \to -\infty)$ (to make h_t bounded), which all lead to unbiased estimating equations because of Theorem 8.5. The above estimating equations are then also unbiased because of Lebesgue's dominated convergence theorem [the dominating function is integrable since

$$\begin{aligned} E\int_0^\tau |\log t|\big(dN(t) + \lambda_{\xi,\gamma,\theta}(t)\,dt\big) &= 2E\int_0^\tau |\log t|\lambda_{\xi,\gamma,\theta}(t)\,dt \\ &\le 2\xi\gamma e^{|\theta_1|+|\theta_2|}\int_0^\tau |\log t| t^{\gamma-1}\,dt, \end{aligned}$$

which is finite since $\gamma > 0$].

Under the model for D of Example 5.1,

$$D_\psi(y,t;\overline{Z}_t) = (1-e^\psi)1_{\{\text{treated at } t\}},$$

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

if the patient did not die before time t. In that case these are five unbiased estimating equations for five unknown parameters. If the parameter ψ is of dimension greater than 1, more unbiased estimating equations can be constructed by adding more terms of the form $\alpha f(X_\psi(t), \overline{Z}_{t-})$.

9. $X(t)$ does not predict treatment changes: a martingale result. We show that, under no unmeasured confounding, just as $Y^{(t)}$, $X(t)$ does not predict treatment changes, given past treatment- and covariate history $\overline{Z}_{t-}$. We could hope for that since $X(t) \sim Y^{(t)}$ given $\overline{Z}_t$ (see Section 6). The formal statement is (compare with Assumption 4.3, no unmeasured confounding) the following: the intensity process $\lambda(t)$ of N with respect to $\sigma(\overline{Z}_t)$ is also the intensity process of N with respect to $\sigma(\overline{Z}_t, X(t))^a$. Then $M(t) = N(t) - \int_0^t \lambda(s)\,ds$ is also a martingale with respect to $\sigma(\overline{Z}_t, X(t))^a$. That will be useful later when we study the behavior of estimators $\hat{\theta}$ and $\hat{\psi}$ which are constructed with estimating equations of the form of Theorem 8.5, $P_n \int_0^\tau h_t(X_\psi(t), \overline{Z}_{t-})(dN(t) - \lambda_\theta(t)\,dt) = 0$. For example, we can use the fact that usually $\int_{[0,t]} H(s)\,dM(s)$ is a martingale if M is a martingale and H a predictable process; a sufficient condition for this is that $E\int |H(s)||dM(s)| < \infty$ [see, e.g., Andersen et al. (1993)]. Hence, all estimating equations of the above form which we saw before are in fact martingales for $(\theta,\psi) = (\theta_0,\psi_0)$.

Before going on, we first clarify why $\sigma(\overline{Z}_t, X(t))$ is indeed a filtration. For $s < t$, $X(s)$ is a deterministic (though unknown) function of $(\overline{Z}_{t-}, X(t))$ (i.e., if solutions to the differential equation with D are unique; see, e.g., Theorem A.1 in the Appendix). Similarly, for $s < t$, $X(t)$ is a deterministic function of $(\overline{Z}_{t-}, X(s))$. In the rest of this article we will assume that these functions are measurable functions on $\overline{\mathcal{Z}}_{t-} \times \mathbb{R}$ (sufficient conditions for that are that the infinitesimal shift-function D satisfies regularity Assumption 9.1 below and that for each $\omega \in \Omega$, Z only jumps finitely many times; see Appendix C, Lemma C.1). Thus,

$$\sigma(\overline{Z}_t, X(t)) = \sigma\big(\overline{Z}_t, \big(X(s) : s \le t\big)\big) = \sigma(\overline{Z}_t, X(0)). \tag{15}$$

We will use the filtration $\sigma(\overline{Z}_t, X(t))$ below, keeping in mind that it is indeed a filtration and satisfies equation (15).

In the rest of this section we assume that the infinitesimal shift-function D satisfies the following regularity condition:

ASSUMPTION 9.1 (Regularity of the infinitesimal shift-function D).

(a) (*Continuity between the jump times of* Z). If Z does not jump in (t_1, t_2), then $D(y, t; \overline{Z}_t)$ is continuous in (y, t) on $[t_1, t_2)$ and can be continuously extended to $[t_1, t_2]$.

(b) (*Boundedness*). For each $\omega \in \Omega$, there exists a constant $C(\omega)$ such that $|D(y, t; \overline{Z}_t)| \leq C(\omega)$ for all $t \in [0, \tau]$ and all y.

(c) (*Lipschitz continuity*). For each $\omega \in \Omega$, there exist constants $L_1(\omega)$ and $L_2(\omega)$ with

$$|D(y, t; \overline{Z}_t) - D(z, t; \overline{Z}_t)| \leq L_1(\omega)|y - z|$$

for all $t \in [0, \tau]$ and all y, z and

$$|D(y, t; \overline{Z}_t) - D(y, s; \overline{Z}_s)| \leq L_2(\omega)|t - s|$$

if $s < t$ and Z does not jump in $(s, t]$.

Most regularity conditions on D here are satisfied for the D's from Appendix B [see also Lok (2001, 2004)]. Only the second Lipschitz condition is extra. The Lipschitz conditions are satisfied, for example, if, in between the jump times of Z, D is continuously differentiable with respect to y and t with derivatives which are bounded for every fixed $\omega \in \Omega$.

The next theorem states that M is indeed also a martingale with respect to $\sigma(\overline{Z}_t, X(t))^a$:

THEOREM 9.2. *Suppose that the conditions of Theorem 8.5 hold*: *Assumptions 4.1 (bounded intensity process), 4.2 ($Y^{(\cdot)}$ cadlag), 4.3 (no unmeasured confounding), 8.4 (no instantaneous treatment effect) and for every $t \in [0, \tau]$, $X(t)$ has the same distribution as $Y^{(t)}$ given $\overline{Z}_t$. Suppose, furthermore, that for each $\omega \in \Omega$, Z jumps at most finitely many times, and that D satisfies regularity Condition 9.1. Then the intensity process $\lambda(t)$ of N with respect to $\sigma(\overline{Z}_t)$ is also the intensity process of N with respect to the filtration $\sigma(\overline{Z}_t, X(t))^a$.*

Recall that in Section 6 we already mentioned that, under regularity conditions, $X(t)$ mimics $Y^{(t)}$ in the sense that it has the same distribution as $Y^{(t)}$ given $\overline{Z}_t$.

Under local rank preservation (see Section 7), $X(t) = Y^{(t)}$ for each t. In that case Theorem 9.2 would be the same as the Assumption of no unmeasured confounding 4.3. However, as argued in Section 7, local rank preservation is generally considered implausible.

PROOF OF THEOREM 9.2. Because of Assumption 4.1, $\Lambda(t) = \int_0^t \lambda(s)\, ds$ is predictable with respect to $\sigma(\overline{Z}_t)$, so then it is also predictable with respect to the larger filtration $\sigma(\overline{Z}_t, X(t))^a$. We still need to prove that M is a martingale with respect to $\sigma(\overline{Z}_t, X(t))^a$. Since a cadlag martingale with respect to some filtration is also a martingale with respect to its usual augmentation [see Rogers and Williams

(1994), Lemma 67.10], it suffices to prove that M is a martingale with respect to $\sigma(\overline{Z}_t, X(t))$. Thus we need to prove that, for $t_2 > t_1$,

$$E[M(t_2) - M(t_1) | \overline{Z}_{t_1}, X(t_1)] = 0.$$

This is not immediate, since we do not have or expect that $(X(t_1), \overline{Z}_{t_2}) \sim (Y^{(t_1)}, \overline{Z}_{t_2})$ if $t_1 < t_2$.

By the definition of conditional expectation, the above is the same as

$$\int_B \big(M(t_2) - M(t_1)\big)\, dP = 0 \tag{16}$$

for all $B \in \sigma(\overline{Z}_{t_1}, X(t_1))$. Because of Theorem 34.1 in Billingsley (1986), it is sufficient to consider B's forming a π-system generating $\sigma(\overline{Z}_{t_1}, X(t_1))$. With σ_1 the σ-algebra on $\overline{Z}_{t_1}$,

$$\{\omega \in \Omega : \overline{Z}_{t_1} \in A \text{ and } X(t_1) \in (x_1, x_2) : A \in \sigma_1 \text{ and } x_1 < x_2 \in \mathbb{R}\}$$

is such a π-system: it is closed under the formation of finite intersections and generates $\sigma(\overline{Z}_{t_1}, X(t_1))$. Therefore, we only consider B's of this form. We prove (16) for any $B = \{\overline{Z}_{t_1} \in A\} \cap \{X(t_1) \in (x_1, x_2)\}$. Let $1^{(n)}_{(x_1,x_2)}$ be any approximation of $1_{(x_1,x_2)}$ which is continuous for every fixed n, with $1^{(n)}_{(x_1,x_2)}(x) \to 1_{(x_1,x_2)}(x)$ for every x as $n \to \infty$ and $|1^{(n)}_{(x_1,x_2)}| \le 1$ for all x and n. Then

$$\begin{aligned}
&\int_B \big(M(t_2) - M(t_1)\big)\, dP \\
&\quad = E\big(1_B \cdot \big(M(t_2) - M(t_1)\big)\big) \\
&\quad = E\bigg(1_A(\overline{Z}_{t_1}) 1_{(x_1,x_2)}(X(t_1)) \int_{(t_1,t_2]} dM(t)\bigg) \\
&\quad = E \int 1_{(t_1,t_2]}(t)\, 1_A(\overline{Z}_{t_1}) 1_{(x_1,x_2)}(X(t_1))\, dM(t) \\
&\quad = E \int 1_{(t_1,t_2]}(t)\, 1_A(\overline{Z}_{t_1}) \lim_{n\to\infty} 1^{(n)}_{(x_1,x_2)}(X(t_1))\, dM(t) \\
&\quad = E \lim_{n\to\infty} \int 1_{(t_1,t_2]}(t)\, 1_A(\overline{Z}_{t_1}) 1^{(n)}_{(x_1,x_2)}(X(t_1)) \big(dN(t) - \lambda(t)\, dt\big) \\
&\quad = \lim_{n\to\infty} E \int 1_{(t_1,t_2]}(t)\, 1_A(\overline{Z}_{t_1}) 1^{(n)}_{(x_1,x_2)}(X(t_1))\, dM(t).
\end{aligned}$$

The last two equalities follow from Lebesgue's dominated convergence theorem [the prior to last equality since, for $\omega \in \Omega$ fixed, the integral is bounded since N is finite and λ is bounded; the last equality since the integrals are all bounded by $N(\tau) + \int_0^\tau \lambda(t)\, dt$, whose expectation is bounded by 2τ times the upper bound of λ]. Equation (16) and the result of the theorem would follow from Theorem 8.5 if

$$1_{(t_1,t_2]}(t)\, 1_A(\overline{Z}_{t_1}) 1^{(n)}_{(x_1,x_2)}(X(t_1)) = h^{(n)}_t(X(t), \overline{Z}_{t-})$$

for some $h_t^{(n)}$ from $\mathbb{R} \times \overline{\mathcal{Z}}_{t-} \to \mathbb{R}$ satisfying Restriction 8.1 for each fixed n. In principle, this seems possible, since $X(t_1)$ is a function of $X(t)$ and $\overline{Z}_{t-}$ for every $t > t_1$.

Indeed, under the conditions above on D and Z, it is possible to find such an $h_t^{(n)}$, as follows. Write $x(\cdot; t_0, x_0)$ for the solution of the differential equation

$$x'(t) = D(x(t), t; \overline{Z}_t)$$

with (final or initial, depending on t) condition $x(t_0) = x_0$. Existence and uniqueness of $x(\cdot; t_0, x_0)$ on $[0, \tau]$ for every fixed $\omega \in \Omega$ follows from Theorem A.1 in Appendix A. In this notation,

$$\begin{aligned} &1_{(t_1,t_2]}(t)\, 1_A(\overline{Z}_{t_1}) 1^{(n)}_{(x_1,x_2)}(X(t_1)) \\ &\quad = 1_{(t_1,t_2]}(t)\, 1_A(\overline{Z}_{t_1}) 1^{(n)}_{(x_1,x_2)}(x(t_1; t, X(t))) = h_t^{(n)}(X(t), \overline{Z}_{t-}) \end{aligned}$$

with

$$h_t^{(n)}(y, \overline{Z}_{t-}) = 1_{(t_1,t_2]}(t)\, 1_A(\overline{Z}_{t_1}) 1^{(n)}_{(x_1,x_2)}(x(t_1; t, y)). \tag{17}$$

We have to show that (17) satisfies Restriction 8.1. First we show that, for t fixed, $h_t^{(n)} : \mathbb{R} \times \overline{\mathcal{Z}}_{t-} \to \mathbb{R}$ is measurable. From (17) we see that this is the case if $x(t_1; t, \cdot) : \mathbb{R} \times \overline{\mathcal{Z}}_{t-} \to \mathbb{R}$ is measurable, which follows immediately from Lemma C.1. Restriction 8.1(a) is immediate, since $h_t^{(n)}$ is bounded by 1. For Restriction 8.1(b), we have to prove that, for all $\omega \in \Omega$, $h_t^{(n)}(y, \overline{Z}_{t-}(\omega)) \to h_{t_0}^{(n)}(y_0, \overline{Z}_{t_0-}(\omega))$ when $y \to y_0$ and $t \uparrow t_0$. Fix $\omega \in \Omega$. We consider three different kinds of t_0. If $t_0 \le t_1$ and $t \uparrow t_0$, $h_t^{(n)}() = 0 = h_{t_0}^{(n)}(\cdot)$, so that the convergence follows immediately. If $t_0 > t_2$ and $t \uparrow t_0$, eventually $h_t^{(n)}(\cdot) = 0 = h_{t_0}^{(n)}(\cdot)$, so that the convergence also follows immediately. If $t_0 \in (t_1, t_2]$, convergence of the first two factors is immediate. For the last factor, we need differential equation theory. $1^{(n)}_{(x_1,x_2)}$ is continuous. Thus, to prove that the last factor in equation (17) converges, it suffices to show that $x(t_1; t, y) \to x(t_1; t_0, y_0)$ as $t \uparrow t_0$ and $y \to y_0$.

Fix $\omega \in \Omega$. For t close enough to t_0, we compare the solution of the differential equation with final condition y at t with the solution of the differential equation with final condition y_0 at t_0; we look at the value of the solution at the time point t_1 before both t and t_0. First, notice that because of existence and uniqueness of solutions (Theorem A.1), the solution of the differential equation with final condition y at t takes a unique value $\tilde{y} = x(t_0; t, y)$ at t_0. Since x is differentiable with respect to its first argument with derivative D and D is bounded by $C(\omega)$ [Assumption 9.1(b)], $\tilde{y}$ is not far from y if t is not far from t_0:

$$|\tilde{y} - y| = |x(t_0; t, y) - y| \le C(\omega)|t - t_0|. \tag{18}$$

Next, notice that, again because of existence and uniqueness of solutions, the value at t_1 of the solution of the differential equation with final condition y at t is the

same as the value at t_1 of the solution of the differential equation with final condition $\tilde{y} = x(t_0; t, y)$ at t_0. This observation implies that

$$
\begin{aligned}
|x(t_1; t, y) - x(t_1; t_0, y_0)| &= |x(t_1; t_0, \tilde{y}) - x(t_1; t_0, y_0)| \\
&\le e^{L_1(\omega)|t_1 - t_0|} |\tilde{y} - y_0| \\
&\le e^{L_1(\omega)|t_1 - t_0|} (|\tilde{y} - y| + |y - y_0|) \\
&\le e^{L_1(\omega)|t_1 - t_0|} \big(C(\omega)|t - t_0| + |y - y_0|\big).
\end{aligned} \tag{19}
$$

For the first inequality, we use Corollary A.3 and Assumption 9.1 [notice that possible jumps of D at the jump times of Z do not matter here since one can split up the interval, so if, e.g., there is just one jump at $\tilde{t} \in (t_1, t_0]$, one gets a factor

$$
e^{L_1(\omega)|t_1 - \tilde{t}|} \cdot e^{L_1(\omega)|\tilde{t} - t_0|} = e^{L_1(\omega)|t_1 - t_0|},
$$

etc. (a formal proof can be given with induction since, with $\omega \in \Omega$ still fixed, there are only finitely many jumps of Z)]. For the last inequality, we use equation (18). If $y \to y_0$ and $t \uparrow t_0$, the bound in equation (19) converges to 0 for every fixed $\omega \in \Omega$. Thus, indeed, if $y \to y_0$ and $t \uparrow t_0$, $x(t_1; t, y)$ converges to $x(t_1; t_0, y_0)$. This finishes the proof. □

10. Consistency and asymptotic normality. The estimating equations for (θ, ψ) from Section 8 were all of the form $P_n g_{\theta,\psi}(Y, \overline{Z}) = 0$. In the current section we choose the dimension of g the same as the dimension of (θ, ψ). Estimating equations of this form are well known. Theorem 10.2 below is an example of asymptotic theory in the setting of this article, with conditions in terms of h and the intensity process λ. Notice, however, that these conditions are in fact stronger than necessary. For more theory about these types of estimating equations and less restrictive conditions, see Van der Vaart (1998), Chapter 5. In particular, conditions could be weakened by considering the estimating equations as a whole instead of looking at h and λ separately (see, e.g., Example 10.5).

We only consider smooth $h_t^{\theta,\psi}$:

RESTRICTION 10.1. The functions $h_t^{\theta,\psi} : \mathbb{R} \times \overline{\mathbb{Z}}_t \to \mathbb{R}^m$ are measurable and:

(a) Every component of $h_t^{\theta_0,\psi_0}$ satisfies Restriction 8.1.

(b) $h_t^{\theta,\psi}(y, \overline{Z}_{t-})$ is bounded by a constant C_1 not depending on θ, ψ, t, y and $\omega \in \Omega$.

(c) For each $t \in [0, \tau]$ and $\omega \in \Omega$, $(\theta, \psi, y) \to h_t^{\theta,\psi}(y, \overline{Z}_{t-})$ is continuous.

(d) There exists a neighborhood of (θ_0, ψ_0) such that, for each $t \in [0, \tau]$ and $\omega \in \Omega$, $h_t^{\theta,\psi}(y, \overline{Z}_{t-})$ is continuously differentiable with respect to θ, ψ and y, and these derivatives are all bounded by a constant C_2 not depending on θ, ψ, t, y and $\omega \in \Omega$.

(e) Every component of $\frac{\partial}{\partial(\theta,\psi)}|_{(\theta,\psi)=(\theta_0,\psi_0)} h_t^{\theta,\psi}(y, \overline{Z}_{t-})$ satisfies Restriction 8.1.

THEOREM 10.2 (Consistency and asymptotic normality). *Suppose that Assumptions* 4.1 (*bounded intensity process*), 4.2 [$Y^{(\cdot)}$ *cadlag*], 4.3 (*no unmeasured confounding*) *and* 8.4 (*no instantaneous treatment effect*) *are satisfied. Suppose also that, for every* $t \in [0, \tau]$, $X(t)$ *has the same distribution as* $Y^{(t)}$ *given* $\overline{Z}_t$. *From Theorem* 8.5 *we know that, for h satisfying Restriction* 10.1(a), (θ_0, ψ_0) *is a zero of*

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

Suppose now that (θ_0, ψ_0) *is the only zero. Suppose, furthermore, that we know that* $(\theta_0, \psi_0) \in (\Theta, \Psi)$ *with* (Θ, Ψ) *compact, that* $\theta \to \lambda_\theta(t)$ *is continuous for each t and bounded by a constant* C_3 *which does not depend on* (ω, t, θ), *and that* $\psi \to X_\psi(t)$ *is continuous for each t. Then any sequence of* (*almost*) *zeros* $(\hat{\theta}, \hat{\psi})$ *of*

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

that is, any sequence of estimators $(\hat{\theta}, \hat{\psi})$ *such that* $\Psi_n(\hat{\theta}, \hat{\psi})$ *converges in probability to zero, is a consistent estimator for* (θ_0, ψ_0) *for each* h_t *satisfying Restriction* 10.1(a)–(c).

Suppose, moreover, that $\theta \to \lambda_\theta(t)$ *is differentiable with the respect to* θ *with derivative bounded by a constant* C_4 *in a neighborhood of* θ_0, *and* $\psi \to X_\psi(t)$ *is differentiable with respect to* ψ *with the derivative bounded by a constant* C_5 *in a neighborhood of* ψ_0. *Then for each h satisfying Restriction* 10.1 *there is a neighborhood of* (θ_0, ψ_0) *such that* $E\int_0^\tau h_t^{\theta,\psi}(X_\psi(t), \overline{Z}_{t-})(dN(t) - \lambda_\theta(t)\,dt)$ *is continuously differentiable with respect to* (θ, ψ). *Suppose, moreover, that the matrix*

$$V_0 = E\bigg(\frac{\partial}{\partial(\theta,\psi)}\bigg|_{(\theta,\psi)=(\theta_0,\psi_0)} \int_0^\tau h_t^{\theta,\psi}(X_\psi(t), \overline{Z}_{t-})\big(dN(t) - \lambda_\theta(t)\,dt\big)\bigg)$$

is nonsingular. Then there exists a sequence of (*almost*) *zeros* $(\hat{\theta}, \hat{\psi})$ *to* $\Psi_n(\theta, \psi)$. *Furthermore, any such sequence is asymptotically normal*:

$$\sqrt{n}\big((\hat{\theta} \quad \hat{\psi})^\top - (\theta_0 \quad \psi_0)^\top\big) \rightsquigarrow \mathcal{N}(0, V_0^{-1} W_0 (V_0^{-1})^\top) \tag{20}$$

with V_0 *the matrix above and, with* $a^{\otimes 2} = aa^\top$,

$$W_0 = E\bigg(\bigg(\int_0^\tau h_t^{\theta_0,\psi_0}(X_{\psi_0}(t), \overline{Z}_{t-})\big(dN(t) - \lambda_{\theta_0}(t)\big)\bigg)^{\otimes 2}\bigg).$$

PROOF. Consistency follows from Theorem 5.9 of Van der Vaart (1998). Repeatedly applying Lebesgue's dominated convergence theorem shows that our conditions imply the conditions of the first and second paragraph after this Theorem 5.9.

The existence of (almost) zeros follows from Van der Vaart and Wellner (1996), Section 3.9, Problem 9, whose solution is practically given by the hint below it. This Problem 9 states that if $f:\Theta\times\Psi\to\mathbb{R}^d$ is a homeomorphism of a neighborhood of $(\theta_0,\psi_0)\in\mathbb{R}^d$ onto a neighborhood of $0\in\mathbb{R}^d$, then every continuous $f:\Theta\times\Psi\to\mathbb{R}^d$ for which $\sup_{(\theta,\psi)\in\Theta\times\Psi}\|f(\theta,\psi)-g(\theta,\psi)\|$ is sufficiently small has at least one zero. In our case $g(\theta,\psi)=E\int_0^\tau h_t^{\theta,\psi}(X_\psi(t),\overline{Z}_{t-})(dN(t)-\lambda_\theta(t))$ is continuously differentiable in a neighborhood of (θ_0,ψ_0) by Restriction 10.1(d) and the assumptions on λ_θ and $X_\psi(t)$, under which differentiation and integration can be exchanged (twice). The derivative of this $g(\theta,\psi)$ at (θ_0,ψ_0) is nonsingular by assumption, and hence, g is a homeomorphism of a neighborhood of $(\theta_0,\psi_0)\in\mathbb{R}^d$ onto a neighborhood of $0\in\mathbb{R}^d$. $\Psi_n(\theta,\psi)$ is continuous in (θ,ψ) and close enough to $g(\theta,\psi)$ for large n with probability approaching 1 because of the second paragraph below Theorem 5.9 in Van der Vaart (1998). Hence, $\Psi_n(\theta,\psi)$ has a zero with probability approaching 1.

Asymptotic normality follows from Theorem 5.21 of Van der Vaart (1998), as follows. Define $g_{\theta,\psi}(Y,\overline{Z}_t)=\int_0^\tau h_t^{\theta,\psi}(X_\psi(t),\overline{Z}_{t-})(dN(t)-\lambda_\theta(t)\,dt)$, which is continuously differentiable with respect to (θ,ψ) in a neighborhood U of (θ_0,ψ_0) under our conditions. Making U smaller so that all boundedness conditions hold on U, we define $\dot g(Y,\overline{Z}_t)=\sup_{(\theta,\psi)\in U}\|\frac{\partial}{\partial(\theta,\psi)}g_{\theta,\psi}(Y,\overline{Z}_t)\|$, which is bounded by $(C_2+C_2C_5+C_2)(N(\tau)+C_3\tau)+C_1C_4\tau$, a constant plus a constant times $N(\tau)$. This $\dot g(Y,\overline{Z}_t)$ is square integrable since $N(\tau)$ is square integrable: it is well known that counting processes with bounded intensity processes are square integrable [it follows from, e.g., Proposition II.4.1 of Andersen et al. (1993)]. For the same reason, $E\|g_{\theta,\psi}(Y,\overline{Z}_t)\|^2<\infty$. The remaining conditions of Theorem 5.21 from Van der Vaart (1998) were checked before, so, indeed, $(\hat\theta,\hat\psi)$ is asymptotically normal with asymptotic covariance matrix (20). □

The asymptotic variance (20) is often estimated by replacing (θ_0,ψ_0) by their estimates and E by P_n. Thus, confidence intervals for ψ_0 can be constructed. Also, tests for whether ψ_0 has a specific value can be constructed that way. For more about testing, see Section 11.

One can often simplify the expression for the asymptotic variance in equation (20) using Corollary 10.4 below. We use the following lemma:

LEMMA 10.3. *Suppose that the conditions of Theorem 9.2 hold*: *Assumptions 4.1 (bounded intensity process), 4.2 [$Y^{(\cdot)}$ cadlag], 4.3 (no unmeasured confounding), 8.4 (no instantaneous treatment effect), for every $t\in[0,\tau]$, $X(t)$ has the same distribution as $Y^{(t)}$ given $\overline{Z}_t$ (see Section 6), for each $\omega\in\Omega$, Z jumps at most*

finitely many times, and D satisfies regularity Condition 9.1. Write $a \otimes b = ab^\top$ and $a^{\otimes 2} = aa^\top$. Then for $h_t : \mathbb{R} \times \overline{\mathcal{Z}}_t \to \mathbb{R}^m$ every component of which satisfies Restriction 8.1,

$$E\left(\left(\int_0^\tau h_t^{\theta_0,\psi_0}(X(t), \overline{Z}_{t-})\, dM(t)\right)^{\otimes 2}\right) = E\int_0^\tau h_t^{\theta_0,\psi_0}(X(t), \overline{Z}_{t-})^{\otimes 2}\lambda(t)\, dt.$$

If, furthermore, λ_θ is a correctly specified model for λ such that $\frac{\partial}{\partial\theta}\lambda_\theta$ exists and D_ψ is a correctly specified model for D such that, for each t, $X_\psi(t)$ is differentiable with respect to ψ at $\psi = \psi_0$, then, for $h_t^{\theta,\psi}$ satisfying Restriction 10.1,

$$\begin{aligned} &E\frac{\partial}{\partial\theta}\bigg|_{\theta=\theta_0}\int_0^\tau h_t^{\theta,\psi_0}(X(t), \overline{Z}_{t-})\big(dN(t) - \lambda_\theta(t)\, dt\big) \\ &\quad = -E\int_0^\tau h_t^{\theta_0,\psi_0}(X(t), \overline{Z}_{t-}) \cdot \frac{\partial}{\partial\theta}\bigg|_{\theta=\theta_0}\lambda_\theta(t)\, dt \end{aligned}$$

if the left- or right-hand side exists and

$$\begin{aligned} &E\frac{\partial}{\partial\psi}\bigg|_{\psi=\psi_0}\int_0^\tau h_t^{\theta_0,\psi}(X_\psi(t), \overline{Z}_{t-})\, dM(t) \\ &\quad = E\int_0^\tau \tilde{h}(X(t), \overline{Z}_{t-}) \otimes \frac{\partial}{\partial\psi}\bigg|_{\psi=\psi_0} X_\psi(t)\, dM(t) \end{aligned}$$

if the left- or right-hand side exists, where $\tilde{h}(y, \overline{Z}_{t-}) = \frac{\partial}{\partial y}h_t^{\theta_0,\psi_0}(y, \overline{Z}_{t-})$.

PROOF. For the first statement, we use counting process theory from Andersen et al. (1993), Chapter 2. If M_1 is a martingale, $\langle M_1\rangle$ (if it exists) is defined as a predictable process such that $M_1^2 - \langle M_1\rangle$ is a (local) martingale. If M_2 is another martingale, $\langle M_1, M_2\rangle$ (if it exists) is defined as a predictable process such that $M_1M_2 - \langle M_2, M_2\rangle$ is a (local) martingale. $\langle M_1\rangle$ is called the predictable variation process of M_1 and $\langle M_1, M_2\rangle$ is called the predictable covariation process of M_1 and M_2. For vector-valued M_1, $\langle M_1\rangle$ is defined as a predictable process such that $M_1^{\otimes 2} - \langle M_1\rangle$ is a (local) martingale. Hence, it is a matrix with $\langle M_{1i}, M_{1j}\rangle$ at the ith row, jth column.

As shown in Theorem 9.2, $M(t) = N(t) - \int_0^t \lambda(s)\, ds$ is a martingale with respect to the filtration $\sigma(\overline{Z}_t, X(t))^a$. Counting process martingales like this have compensators: $\langle M(t)\rangle = \int_0^t \lambda(s)\, ds$ [see, e.g., Proposition II.4.1 in Andersen et al. (1993)]. Moreover, if H_1 and H_2 are (locally) bounded $\sigma(\overline{Z}_t, X(t))^a$-predictable processes with values in $\mathbb{R}$, then $\langle\int_0^t H_1(s)\, dM(s), \int_0^t H_2(s)\, dM(s)\rangle$ exists and

$$\left\langle\int_0^t H_1(s)\, dM(s), \int_0^t H_2(s)\, dM(s)\right\rangle = \int_0^t H_1(s)H_2(s)\lambda(s)\, ds$$

[Proposition II.4.1 or (2.4.9) in Andersen et al. (1993)]. Because h_t satisfies Restriction 8.1, $h_t(X(t), \overline{Z}_{t-})$ is a bounded $\sigma(\overline{Z}_t, X(t))^a$-predictable process (proof

just as in Lemma 8.2). Therefore, the theory above leads to

$$E\left(\left(\int_0^\tau h_t^{\theta_0,\psi_0}(X(t),\overline{Z}_{t-})\,dM(t)\right)^{\otimes 2}\right) = E\left(\left\langle \int_0^\tau h_t^{\theta_0,\psi_0}(X(t),\overline{Z}_{t-})\,dM(t)\right\rangle\right)$$
$$= E\left(\int_0^\tau h_t^{\theta_0,\psi_0}(X(t),\overline{Z}_{t-})^{\otimes 2}\lambda(t)\,dt\right).$$

For the second statement, notice that, under the conditions of the lemma,

$$\frac{\partial}{\partial\theta}\bigg|_{\theta=\theta_0}\left(\int_0^\tau h_t^{\theta,\psi_0}(X(t),\overline{Z}_{t-})\big(dN(t)-\lambda_\theta(t)\,dt\big)\right)$$
$$= -\int_0^\tau h_t^{\theta_0,\psi_0}(X(t),\overline{Z}_{t-})\otimes\frac{\partial}{\partial\theta}\bigg|_{\theta=\theta_0}\lambda_\theta(t)\,dt$$
$$+\int_0^\tau\left(\frac{\partial}{\partial\theta}\bigg|_{\theta=\theta_0}h_t^{\theta,\psi_0}\right)(X(t),\overline{Z}_{t-})\big(dN(t)-\lambda(t)\,dt\big),$$

and the expectation of the second term here is equal to zero because of Theorem 8.5.

For the third statement, notice that, under the conditions of the lemma,

$$\frac{\partial}{\partial\psi}\bigg|_{\psi=\psi_0}\left(\int_0^\tau h_t^{\theta_0,\psi}(X_\psi(t),\overline{Z}_{t-})\big(dN(t)-\lambda_\theta(t)\,dt\big)\right)$$
$$= \int_0^\tau\left(\tilde{h}_t(X(t),\overline{Z}_{t-})\otimes\frac{\partial}{\partial\psi}\bigg|_{\psi=\psi_0}X_\psi(t)\right)\big(dN(t)-\lambda(t)\,dt\big)$$
$$+\int_0^\tau\left(\frac{\partial}{\partial\psi}\bigg|_{\psi=\psi_0}h_t^{\theta_0,\psi}\right)(X(t),\overline{Z}_{t-})\big(dN(t)-\lambda(t)\,dt\big)$$

because of the chain rule, and the expectation of the second term is equal to zero because of Theorem 8.5. □

This lemma simplifies the asymptotic variance formula of the estimators in equation (20):

COROLLARY 10.4 (Asymptotic variance). *Suppose that the conditions of Theorem 9.2 hold*: *Assumptions 4.1 (bounded intensity process), 4.2 [$Y^{(\cdot)}$ cadlag], 4.3 (no unmeasured confounding), 8.4 (no instantaneous treatment effect), for every $t\in[0,\tau]$, $X(t)$ has the same distribution as $Y^{(t)}$ given $\overline{Z}_t$ (see Section 6), for each $\omega\in\Omega$, Z jumps at most finitely many times, and D satisfies regularity Condition 9.1. Suppose also that λ_θ is a correctly specified model for λ such that $\frac{\partial}{\partial\theta}\lambda_\theta$ exists and that D_ψ is a correctly specified model for D such that, for each t, $X_\psi(t)$ is differentiable with respect to ψ at $\psi=\psi_0$. Then if*

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

and $h_t^{\theta,\psi}$ *satisfies Restriction* 10.1, *the asymptotic variance* (20) *is equal to* $V_0^{-1} W_0 V_0^{-1^\top}$ *with*

$$W_0 = E\left(\int_0^\tau h_t^{\theta_0,\psi_0}(X(t), \overline{Z}_{t-})^{\otimes 2} \lambda(t)\, dt\right)$$

and $V_0 = (V_{0\theta} V_{0\psi})$ *with*

$$V_{0\theta} = -E\left(\int_0^\tau \left(h_t^{\theta_0,\psi_0}(X(t), \overline{Z}_{t-}) \otimes \frac{\partial}{\partial\theta}\bigg|_{\theta=\theta_0} \lambda_\theta(t)\right) dt\right)$$

and, with $\tilde{h}(y, \overline{Z}_{t-}) = \frac{\partial}{\partial y} h_t^{\theta_0,\psi_0}(y, \overline{Z}_{t-})$,

$$V_{0\psi} = E\left(\int_0^\tau \left(\tilde{h}(X(t), \overline{Z}_{t-}) \otimes \frac{\partial}{\partial\psi}\bigg|_{\psi=\psi_0} X_\psi(t)\right)\big(dN(t) - \lambda(t)\, dt\big)\right).$$

We conclude this section with an example to see the machinery work in practice. Notice that the boundedness conditions of Theorem 10.2 are somewhat too restrictive for this example, but that the results hold true under these weaker restrictions, too.

LEMMA 10.5 (Survival of AIDS patients and the Weibull proportional hazards model). *Consider the setting of Example* 8.6, *and suppose that the assumptions of Section* 9 *are satisfied. In Example* 8.6,

$$\lambda_{\xi,\gamma,\theta}(t) = 1_{\{\text{at risk at } t\}} \xi\gamma t^{\gamma-1} e^{\theta_1 I_{\text{AZT}} + \theta_2 I_{\text{PCP}}(t)},$$

where at risk means at risk for initiation of prophylaxis treatment. $X_\psi(t)$ *is the solution to the differential equation* $X'_\psi(t) = D_\psi(X_\psi(t), t; \overline{Z}_t)$ *with final condition* $X_\psi(\tau) = Y$ *and* $D_\psi(y, t; \overline{Z}_t) = (1 - e^\psi) 1_{\{\text{treatedat}t\}}$, *so*

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

In Example 8.6 *we already saw that* $(\xi_0, \gamma_0, \theta_0, \psi_0)$ *is a zero of*

$$E\left(\int_0^\tau \begin{pmatrix} \frac{1}{\xi} & \frac{1}{\gamma} + \log t & I_{\text{AZT}} & I_{\text{PCP}}(t) & X_\psi(t) \end{pmatrix}^\top \big(dN(t) - \lambda_{\xi,\gamma,\theta}(t)\, dt\big)\right).$$

Suppose now that $(\xi_0, \gamma_0, \theta_0, \psi_0)$ *is the only zero. Suppose, furthermore, that the survival time* Y *takes values in a compact space* $[0, y_0] \subset \mathbb{R}$ *and that we know that* $(\xi_0, \gamma_0, \theta_0, \psi_0) \in \Xi \times \Gamma \times \Theta \times \Psi$, *with* $\Xi \subset (0, \infty)$, $\Gamma \subset (0, \infty)$, Θ *and* Ψ *all four compact* (*note that this implies that* ξ *and* γ *are bounded away from* 0). *Then any sequence of* (*almost*) *zeros* $(\hat{\xi}, \hat{\gamma}, \hat{\theta}, \hat{\psi})$ *of*

$$\begin{aligned} &\Psi_n(\xi, \gamma, \theta, \psi) \\ &\quad = P_n \int_0^\tau \begin{pmatrix} \frac{1}{\xi} & \frac{1}{\gamma} + \log t & I_{\text{AZT}} & I_{\text{PCP}}(t) & X_\psi(t) \end{pmatrix}^\top \\ &\qquad \times \big(dN(t) - \lambda_{\xi,\gamma,\theta}(t)\, dt\big), \end{aligned} \tag{21}$$

that is, any sequence of estimators $(\hat{\xi}, \hat{\gamma}, \hat{\theta}, \hat{\psi})$ *such that* $\Psi_n(\hat{\xi}, \hat{\gamma}, \hat{\theta}, \hat{\psi})$ *converges in probability to zero, is a consistent estimator for* $(\xi_0, \gamma_0, \theta_0, \psi_0)$. *Moreover,* $V_0 = (V_{0\theta}\, V_{0\psi})$ *as in Corollary 10.4 exists, and*

$$V_{0\theta} = -E\int_0^\tau \begin{pmatrix} \frac{1}{\xi_0} & \frac{1}{\gamma_0} + \log t & I_{\mathrm{AZT}} & I_{\mathrm{PCP}}(t) & X(t) \end{pmatrix}^\top \times \begin{pmatrix} \frac{1}{\xi_0} & \frac{1}{\gamma_0} + \log t & I_{\mathrm{AZT}} & I_{\mathrm{PCP}}(t) \end{pmatrix} \lambda(t)\, dt$$

and $V_{0\psi}$ *is a five-dimensional vector with zeros in the first four positions and*

$$E\int_0^\tau \left(\frac{\partial}{\partial \psi}\bigg|_{\psi=\psi_0} X_\psi(t) \right) \big(dN(t) - \lambda(t)\, dt\big)$$

in the fifth, with

$$\frac{\partial}{\partial \psi}\bigg|_{\psi=\psi_0} X_\psi(t) = \int_{t\wedge Y}^{Y} e^{\psi_0}\, 1_{\{\text{treated at } s\}}\, ds.$$

If this V_0 *is a nonsingular matrix, then there exists a sequence of (almost) zeros* $(\hat{\xi}, \hat{\gamma}, \hat{\theta}, \hat{\psi})$ *of (21). Furthermore, any such sequence is asymptotically normal*:

$$\sqrt{n}\big((\hat{\xi} \quad \hat{\gamma} \quad \hat{\theta} \quad \hat{\psi})^\top - (\xi_0 \quad \gamma_0 \quad \theta_0 \quad \psi_0)^\top\big) \rightsquigarrow \mathcal{N}\big(0, V_0^{-1} W_0 (V_0^{-1})^\top\big),$$

with V_0 *the matrix above and*

$$W_0 = E\int_0^\tau \begin{pmatrix} \frac{1}{\xi_0} & \frac{1}{\gamma_0} + \log t & I_{\mathrm{AZT}} & I_{\mathrm{PCP}}(t) & X(t) \end{pmatrix}^{\otimes 2} \lambda(t)\, dt.$$

Moreover, $\hat{\theta}$ *and* $\hat{\psi}$ *are asymptotically independent.*

The asymptotic independence here turns out to be no coincidence; see Lok (2001, 2007).

PROOF OF LEMMA 10.5. The findings here are similar to the findings in Theorem 10.2, but the boundedness conditions fail to hold here. Consistency follows from Van der Vaart (1998), Theorem 5.9. Existence of a sequence of (almost) zeros follows from Van der Vaart and Wellner (1996), Section 3.9, Problem 9, whose solution is practically given by the hint below it. Asymptotic normality follows from Van der Vaart (1998), Theorem 5.21. The asymptotic variance equals $V_0^{-1} W_0 (V_0^{-1})^\top$ because of Corollary 10.4 (λ and h are not bounded here, but we can restrict the interval to $[\varepsilon, \tau]$ for $\varepsilon > 0$ and let $\varepsilon \downarrow 0$). We leave checking the conditions of these theorems to the reader [or see Lok (2001), Section 7.7].

Asymptotic independence of $\hat{\theta}$ and $\hat{\psi}$ follows by direct calculation, after noticing that

$$V_0 = \begin{pmatrix} -B & \mathbf{0} \\ -C & A_{\psi_0} \end{pmatrix} \Longrightarrow V_0^{-1} = \begin{pmatrix} -B^{-1} & \mathbf{0} \\ -A_{\psi_0}^{-1} C B^{-1} & A_{\psi_0}^{-1} \end{pmatrix},$$

with

$$A_{\psi_0} = E\int_0^\tau \left(\frac{\partial}{\partial\psi}\bigg|_{\psi=\psi_0} X_\psi(t)\right)\big(dN(t) - \lambda(t)\,dt\big),$$

$$B = E\int_0^\tau \begin{pmatrix} \frac{1}{\xi_0} & \frac{1}{\gamma_0} + \log t & I_{\mathrm{AZT}} & I_{\mathrm{PCP}}(t) \end{pmatrix}^{\top^{\otimes 2}} \lambda(t)\,dt$$

and

$$C = E\int_0^\tau X(t)\cdot \begin{pmatrix} \frac{1}{\xi_0} & \frac{1}{\gamma_0} + \log t & I_{\mathrm{AZT}} & I_{\mathrm{PCP}}(t) \end{pmatrix} \lambda(t)\,dt.$$

Define

$$D = E\int_0^\tau X(t)^2\lambda(t)\,dt.$$

This concludes the proof:

$$\begin{aligned} V_0^{-1}W_0(V_0^{-1})^\top &= \begin{pmatrix} -B^{-1} & 0 \\ -A_{\psi_0}^{-1}CB^{-1} & A_{\psi_0}^{-1} \end{pmatrix}\begin{pmatrix} B & C^\top \\ C & D \end{pmatrix} \\ &\quad\times \begin{pmatrix} -B^{-1^\top} & -B^{-1^\top}C^\top A_{\psi_0}^{-1^\top} \\ 0 & A_{\psi_0}^{-1^\top} \end{pmatrix} \\ &= \begin{pmatrix} (B^{-1})^\top & 0 \\ 0 & A_{\psi_0}(CB^{-1}C^\top + D)(A_{\psi_0}^{-1})^\top \end{pmatrix}. \end{aligned}$$

□

11. Test for treatment effect without specifying a model for D. We show that one can often test whether treatment affects the outcome of interest without specifying a model D_ψ for D. This was conjectured, but not proved, in Robins (1998b). If one does not have to specify a model for D in order to test whether treatment affects the outcome, false conclusions caused by misspecification of the model for D can be avoided.

Under the null hypothesis of the no treatment effect, $D \equiv 0$ and $X(t) \equiv Y$ (see Section 5). If there is no unmeasured confounding, $X(t)$ does not predict $N(t)$ given $\overline{Z}_{t-}$ (see Theorem 9.2). Hence, if there is no unmeasured confounding and no treatment effect, adding the *observed outcome* Y to the prediction model for treatment effect should not help the prediction. This idea, presented in Robins (1998b) for the case of local rank preservation (see Section 7), can be proven to be correct as follows.

Technically, the tests in this section are similar to the score test [for more about the score test see, e.g., Cox and Hinkley (1974)]. Suppose that the conditions of Section 9 are satisfied, and that we have a correctly specified parametric model λ_θ for λ. Define

$$g_\theta(Y, \overline{Z}) = \int_0^\tau h_t^\theta(Y, \overline{Z}_{t-})\big(dN(t) - \lambda_\theta(t)\,dt\big),$$

with $h_t^{\theta_0}$ satisfying the regularity condition Restriction 8.1. The key idea of this procedure is that if treatment does not affect the outcome, $D \equiv 0$, so $X(t) \equiv Y$, and $g_{\theta_0}(Y, \overline{Z})$ has expectation zero because of Theorem 8.5. Since θ_0 is unknown, we base the test on the limiting behavior under $D \equiv 0$ of $\sqrt{n} P_n g_{\hat{\theta}}(Y, \overline{Z})$, where $\hat{\theta}$ is an estimator of the nuisance parameter θ_0. We will show that if $D \equiv 0$, $\sqrt{n} P_n g_{\hat{\theta}}(Y, \overline{Z})$ converges to a normal random variable with expectation zero, which leads to a test for whether $D \equiv 0$ in the usual way.

The nuisance parameter θ_0 will be estimated using some set of estimating equations

$$P_n \tilde{g}_\theta(\overline{Z}) = 0$$

with $E\tilde{g}_{\theta_0}(\overline{Z}) = 0$, $E\tilde{g}_\theta(\overline{Z})$ differentiable in θ and $E\tilde{g}_{\theta_0}^2(\overline{Z}) < \infty$. A natural choice would be a maximum (partial) likelihood estimator for θ_0. We suppose throughout this section that the resulting estimator $\hat{\theta}$ is consistent and asymptotically normal with

$$\sqrt{n}(\hat{\theta} - \theta_0) = -\left(\frac{\partial}{\partial \theta}\bigg|_{\theta=\theta_0} E\tilde{g}_\theta(\overline{Z}) \right)^{-1} \sqrt{n} P_n \tilde{g}_{\theta_0}(\overline{Z}) + o_P(1), \tag{22}$$

as will usually follow from, for example, Van der Vaart (1998), Theorem 5.21.

If λ_θ and h_t^θ are sufficiently smooth,

$$\theta \to \sqrt{n} P_n g_\theta(Y, \overline{Z})$$

is differentiable with respect to θ, and a Taylor expansion around θ_0 leads to

$$\sqrt{n} P_n g_{\hat{\theta}}(Y, \overline{Z}) = \sqrt{n} P_n g_{\theta_0}(Y, \overline{Z}) + P_n \dot{g}_{\tilde{\theta}}(Y, \overline{Z}) \sqrt{n}(\hat{\theta} - \theta_0),$$

with $\dot{g}_\theta$ the derivative of g_θ with respect to θ and $\tilde{\theta}$ between θ_0 and $\hat{\theta}$. Since $\hat{\theta}$ converges in probability to θ_0, so does $\tilde{\theta}$. Therefore, usually, $P_n \dot{g}_{\tilde{\theta}}(Y, \overline{Z}) \xrightarrow{P} E\dot{g}_{\theta_0}(Y, \overline{Z})$. Sufficient conditions under which this holds are given in Appendix D, Lemma D.1. Because of (22) and the central limit theorem, $\sqrt{n}(\hat{\theta} - \theta_0)$ converges in distribution. Therefore, an application of Slutzky's lemma leads to

$$\begin{aligned}
\sqrt{n} P_n g_{\hat{\theta}}(Y, \overline{Z}) &= \sqrt{n} P_n g_{\theta_0}(Y, \overline{Z}) + E\dot{g}_{\theta_0}(Y, \overline{Z}) \sqrt{n}(\hat{\theta} - \theta_0) + o_P(1) \\
&= \sqrt{n} P_n g_{\theta_0}(Y, \overline{Z}) - E\dot{g}_{\theta_0}(Y, \overline{Z}) V_0^{-1} \sqrt{n} P_n \tilde{g}_{\theta_0}(\overline{Z}) + o_P(1) \\
&= (-E\dot{g}_{\theta_0}(Y, \overline{Z}) V_0^{-1} \ \mathrm{Id}_{\dim g_\theta}) \begin{pmatrix} \sqrt{n} P_n \tilde{g}_{\theta_0}(\overline{Z}) \\ \sqrt{n} P_n g_{\theta_0}(Y, \overline{Z}) \end{pmatrix} + o_P(1),
\end{aligned}$$

with $V_0 = \frac{\partial}{\partial \theta}|_{\theta=\theta_0} E\tilde{g}_\theta(\overline{Z})$. If $D \equiv 0$, $X(t) \equiv Y$, so that Theorem 8.5 implies that also the expectation of g_{θ_0} is equal to zero. Therefore, the central limit theorem can be applied on the vector with $\sqrt{n}$ on the right-hand side; it converges to a normal random variable with expectation zero. Because of the Continuous Mapping Theorem [see, e.g., Van der Vaart (1998), Chapter 18], $\sqrt{n} P_n g_{\hat{\theta}}(Y, \overline{Z})$ then converges to a normal random variable with expectation zero, too.

Calculation of its limiting covariance matrix is standard [see, e.g., Van der Vaart (1998), Chapter 18]. To save space, we omit that calculation here. If desirable, one can use Theorem 8.5 and Lemma 10.3 to simplify the expression.

Notice that a test for whether $D = D_0$ for any specific D_0 can be constructed in exactly the same way. If we have a correctly specified model D_ψ for D, this thus also leads to a confidence region for ψ_0 in the usual way, using the duality between testing and confidence regions: include those ψ for which the null hypothesis $D = D_\psi$ is not rejected.

12. Discussion and extensions. The proof of consistency and asymptotic normality of the estimators presented in this article applies to continuous-time structural nested models. A similar proof applies to structural nested models in discrete time (when covariates are only measured at finitely many fixed times $0 = \tau_0 < \tau_1 < \cdots < \tau_K < \tau_{K+1} = \tau$, which are the same for all patients and known in advance). Lok, Gill, van der Vaart and Robins (2004) argue without proof that consistency and asymptotic normality should hold for discrete time structural nested models under reasonable assumptions; the proof is completed with the current article, as follows. It is easy to see that in discrete time, $\sum_{\tau_k \le t} P(\Delta N(\tau_k) = 1 | \overline{Z}_{\tau_k -})$ is the compensator of N with respect to $\sigma(\overline{Z}_t)$ [see, e.g., Lok (2001), Section 7.4]. The assumption of no unmeasured confounding can be formalized as

$$P\big(\Delta N(\tau_k) = 1 | \overline{Z}_{\tau_k -}, Y^{(\tau_k)}\big) = P\big(\Delta N(\tau_k) = 1 | \overline{Z}_{\tau_k -}\big).$$

Since $X(\tau_k) \sim Y^{(\tau_k)}$ given $\overline{Z}_{\tau_k}$, this implies that also

$$P\big(\Delta N(\tau_k) = 1 | \overline{Z}_{\tau_k -}, X(\tau_k)\big) = P\big(\Delta N(\tau_k) = 1 | \overline{Z}_{\tau_k -}\big).$$

The discrete-time counterparts of Theorem 8.5 and Theorem 9.2 follow immediately [see Lok (2001)]. Consistency and asymptotic normality follow in the same way as for continuous time models.

The tests for treatment effect in this article can be carried out without specifying a model for treatment effect; that is, no model for D is needed. This is an important feature of the tests because it allows one to avoid false conclusions caused by misspecification of the model for D. In practice, it may be hard to specify a correct parametric model for the infinitesimal shift-function, D. Thus, it is good that specification of a model for D is not needed to test for treatment effect.

The estimators in this article require the correct specification of a model for treatment effect and of a model for prediction of treatment changes. For the discrete-time setting, Robins (2000) has recently proposed estimators which are doubly robust. Doubly robust estimators are consistent and asymptotically normal if (i) the model for prediction of treatment changes (λ in the current article) is correctly specified or if (ii) a regression model of a blipped down outcome [$X_\psi(t)$ in the current article] on past treatment- and covariate history $\overline{Z}_{t-}$ is correctly specified. In any case, the model for treatment effect (D_ψ in the current article) has to be well specified.

In this article estimation started with the specification of a model for the infinitesimal shift-function, D. Interpretation of results may be easier when one starts with a model like (4), $Y^{(t)} - t \sim \int_t^Y e^{\psi 1_{\{\text{treated at } s\}}} ds = e^{\psi} \cdot DUR(t, Y) + 1 \cdot (Y - t - DUR(t, Y))$, given $\overline{Z}_t$. Here, $DUR(t, u)$ is the duration of treatment in the interval (t, u). The main results in the current article apply also to this model. The proofs for Theorems 8.5 and 10.2 do not depend on X being the solution to $X'(t) = D(X(t), t; \overline{Z}_t)$. The proof of Theorem 9.2 does depend on $X'(t) = D(X(t), t; \overline{Z}_t)$, but it simplifies considerably if (4) or (5) is used as a starting point. Let us show this for (4). Define $\tilde{X}(t) = t + \int_t^Y e^{\psi 1_{\{\text{treated at } s\}}} ds$. Using the first part of the proof of Theorem 9.2, for $t_1 < t$,

$$\begin{aligned} &1_{(t_1,t_2]}(t) 1_A(\overline{Z}_{t_1}) 1^{(n)}_{(x_1,x_2)}(\tilde{X}(t_1)) \\ &\quad = 1_{(t_1,t_2]}(t) 1_A(\overline{Z}_{t_1}) 1^{(n)}_{(x_1,x_2)}\big(\tilde{X}(t) + DUR(t_1, t)(-1 + e^{\psi})\big) \\ &\quad = h_t^{(n)}(\tilde{X}(t), \overline{Z}_{t-}), \end{aligned}$$

where $DUR(t_1, t)$ is the duration of treatment in the interval (t_1, t). $h_t^{(n)}$ is measurable if the duration of treatment until t is in included in $Z(t)$. Moreover, if $t \uparrow t_0$ and $y \to y_0$, then $h_t^{(n)}(y, t) \to h_t^{(n)}(y_0, t_0)$. Hence, it follows immediately that $h_t^{(n)}$ satisfies Restriction 8.1, which concludes the proof. Depending on the specific application, it will be more appropriate to start with $Y^{(t)} - t \sim \int_t^Y e^{\psi 1_{\{\text{treated at } s\}}} ds$ given $\overline{Z}_t$ or with a model for D; see, for example, Example 5.3.

In the previous literature [see, e.g., Robins et al. (1992), Mark and Robins (1993), Witteman et al. (1998), Keiding et al. (1999) and Hernán et al. (2005)] applications have been carried out under the assumption of local rank preservation, where $X(t) = Y^{(t)}$ for each t. As pointed out in these papers, and as discussed in Section 7, the assumption of local rank preservation is generally considered implausible. This article relaxes the assumption of local rank preservation in structural nested models. The estimators and tests applied in the previous literature are specific cases of the estimators and tests studied in this article, with the only difference that some of the estimators in the previous literature allow for censoring of Y. Aside from the issue of censoring, this article provides a mathematical foundation behind previous estimators, relaxes the specification of the counterfactual outcomes as deterministic variables, and allows for a distributional interpretation of the estimators.

Robins (1998b) conjectures that one can often use standard software to test whether treatment affects the outcome of interest (without specifying a model D_ψ for D), and to estimate ψ. Lok (2001, 2007) shows that both testing and estimation can also be considered from a partial likelihood point of view. As shown in Lok (2001, 2007), this approach leads to a subclass of the estimators and tests studied in this article which can indeed be calculated with standard software. Example 8.6 is a specific case of that. The possibility to use standard software may be a good reason

to choose these estimators in practice. See Robins (1998b) and Lok (2001, 2007) for a more elaborate discussion.

The approach adopted in the current article leads to a large class of estimators and tests. When treatment and covariates change at finitely many fixed times only, Robins (1993, 1997) proposes, without proof, an optimal procedure for survival and nonsurvival outcomes, respectively. The optimal choice of estimators or tests under the framework of this article is another intriguing topic for future research.

The current article assumes a parametric model λ_θ for the prediction of treatment changes. In practice, applications have often used a semi-parametric Cox model for λ_θ. Lok (2001) shows that specifying λ_θ using a semiparametric Cox model leads to unbiased estimating equations, which just as in this article are martingales for the true parameters. Consistency and asymptotic normality of the resulting estimators for D still remain to be shown and constitute interesting topics for future research.

In many applications, observations are censored. Robins (1998b) and Hernán et al. (2005) have proposed methods to deal with censoring that could potentially be adapted to the results in this paper. For D of the form of Example 5.1, Lok (2007) includes proofs with censoring due to the fact that the study ends, so-called administrative censoring, using ideas from these previous papers.

The estimators in this paper depend on solving a differential equation for each observation. In the examples, these equations are simple enough to be solved analytically. If that is not possible, these equations should be solved numerically. It might be worth investigating how a small contamination of the solution to the differential equation $X(t)$ affects the estimates of treatment effect.

I conclude with a discussion of the assumptions used in this article. The most important assumption in this article is the assumption of no unmeasured confounding (Assumption 4.3). As discussed before, this assumption is valid if all information has been recorded which both (i) predicts treatment decisions and (ii) is an independent risk factor for the outcome of interest. The validity of the assumption of no unmeasured confounding cannot be tested statistically, and depends on the quality of the recorded information. Therefore, it is for subject matter experts to decide about the plausibility of the assumption of no unmeasured confounding. Second, we only estimate the effect of treatment for which a short duration of treatment has only a small effect on the distribution of the outcome of interest. The effect of the treatment on an individual patient may be large, as long as the probability of such an effect is small for any small duration of treatment. Third, the assumption of no instantaneous treatment effect (Assumption 8.4) is also restrictive: it excludes the estimation of the effect of treatments that have instantaneous effects, such as surgery or other point exposures. The remaining assumptions in this article are mostly benign. The assumption that the covariate- and treatment process can be represented by a cadlag process is generally accepted for most medical situations [see, e.g., Andersen et al. (1993)]. The functions h_t and D_ψ can be chosen such that the regularity conditions on these functions are satisfied.

Even if h_t is not bounded, it can often be approximated by bounded functions, and results may follow by a simple application of Lebesgue's dominated convergence theorem [see, e.g., Example 8.6]. The same is true for the boundedness condition (Assumption 4.1) of the intensity process λ. Assumption 4.2 that the counterfactual process $Y^{(t)}$ is cadlag is impossible to verify, but it is a plausible and convenient regularity condition.

APPENDIX A: SOME THEORY ABOUT DIFFERENTIAL EQUATIONS

THEOREM A.1. *Suppose that a function $D(y, t; \overline{Z}_t)$ satisfies the following*:

(a) (continuity between the jump times of Z). *If Z does not jump in (t_1, t_2), then $D(y, t; \overline{Z}_t)$ is continuous in (y, t) on $[t_1, t_2)$ and can be continuously extended to $[t_1, t_2]$.*

(b) (Lipschitz continuity). *For each $\omega \in \Omega$, there exists a constant $L(\omega)$ such that*

$$|D(y, t; \overline{Z}_t) - D(z, t; \overline{Z}_t)| \le L(\omega)|y - z|$$

for all $t \in [0, \tau]$ and all y, z.

Suppose, furthermore, that, for each $\omega \in \Omega$, there are no more than finitely many jump times of Z. Then, for each $t_0 \in [0, \tau]$ and $y_0 \in \mathbb{R}$, there is a unique continuous solution $x(t; t_0, y_0)$ to

$$x'(t) = D(x(t), t; \overline{Z}_t)$$

with boundary condition $x(t_0) = y_0$ and this solution is defined on the whole interval $[0, \tau]$.

This theorem follows from well-known results about differential equations; see, for example, Duistermaat and Eckhaus (1995), Chapter 2.

For the next theorem, we also refer to Duistermaat and Eckhaus (1995), Chapter 2. It is a consequence of Gronwall's lemma.

THEOREM A.2. *Suppose that I is an open or closed interval in $\mathbb{R}$, $f : I \times \mathbb{R}^n \to \mathbb{R}^n$ is continuous and $C : I \to [0, \infty)$ is continuous, and suppose that*

$$\|f(x, y) - f(x, z)\| \le C(x)\|y - z\| \tag{23}$$

for all $x \in I$ and $y, z \in \mathbb{R}^n$. Then, for every $x_0 \in I$ and $y_0 \in \mathbb{R}$, there is a unique solution $y(x)$ of $y'(x) = f(x, y(x))$ with $y(x_0) = y_0$, and this solution is defined for all $x \in I$. If $g : I \times \mathbb{R}^n \to \mathbb{R}^n$ is continuous and $z : I \to \mathbb{R}^n$ is a solution of $z'(x) = g(x, z(x))$, then

$$\begin{aligned}
&\|y(x) - z(x)\| \\
&\quad \le e^{\int_{x_0}^{x} C(\xi)d\xi}\|y(x_0) - z(x_0)\| \\
&\qquad + \int_{x_0}^{x} e^{\int_{\xi}^{x} C(\eta)d\eta}\|f(\xi, z(\xi)) - g(\xi, z(\xi))\|\, d\xi
\end{aligned}$$

for all $x, x_0 \in I$ with $x_0 \le x$.

In Duistermaat and Eckhaus (1995) the interval is always an open interval, but as is generally known, this can be overcome by extending both f and g outside the closed interval I by taking the values at the boundary of I. This preserves the Lipschitz- and continuity conditions. Existence and uniqueness on all of finitely many intervals implies global existence and uniqueness.

We have a differential equation with end condition at τ, so we are interested in x, x_0 with $x \le x_0$:

COROLLARY A.3. *Suppose that the conditions of Theorem A.2 are satisfied. Then, for every $x_0 \in I$ and $y_0 \in \mathbb{R}^n$, there is a unique solution $y(x)$ of $y'(x) = f(x, y(x))$ with $y(x_0) = y_0$, and this solution is defined for all $x \in I$. If $g : I \times \mathbb{R}^n \to \mathbb{R}^n$ is continuous and $z : I \to \mathbb{R}^n$ is a solution of $z'(x) = g(x, z(x))$, then*

$$\begin{aligned}\|y(x) - z(x)\| &\le e^{\int_x^{x_0} C(s)ds} \|y(x_0) - z(x_0)\| \\ &\quad + \int_x^{x_0} e^{\int_x^s C(\eta)d\eta} \|f(s, z(s)) - g(s, z(s))\|\, ds\end{aligned}$$

for all x, x_0 with $x \le x_0$.

PROOF. Put $\tilde{y}(t) = y(x_0 - t)$. Then

$$\tilde{y}'(t) = -y'(x_0 - t) = -f\big(x_0 - t, y(x_0 - t)\big) = \tilde{f}(t, \tilde{y}(t)),$$

where $\tilde{f}(t, y) = -f(x_0 - t, y)$. Thus, $\tilde{y}(t) = y(x_0 - t)$ is a solution of the differential equation $\tilde{y}'(t) = \tilde{f}(t, \tilde{y}(t))$ with boundary condition $\tilde{y}(0) = y(x_0) = y_0$. Define also $\tilde{z}(t) = z(x_0 - t)$. Applying Theorem A.2 on $\tilde{y}$ concludes the proof, as follows:

$$\begin{aligned}\|y(x) - z(x)\| &= \big\|y\big(x_0 - (x_0 - x)\big) - z\big(x_0 - (x_0 - x)\big)\big\| \\ &= \|\tilde{y}(x_0 - x) - \tilde{z}(x_0 - x)\| \\ &= \|\tilde{y}(t) - \tilde{z}(t)\|\end{aligned}$$

with $t = x_0 - x \ge 0$. Notice that, because of equation (23),

$$\|\tilde{f}(t, y) - \tilde{f}(t, z)\| \le C(x_0 - t)\|y - z\| =: \tilde{C}(t)\|y - z\|,$$

with $\tilde{C}(t) = C(x_0 - t)$. Hence, Theorem A.2 implies that

$$\begin{aligned}\|y(x) - z(x)\| &\le e^{\int_0^t \tilde{C}(\xi)d\xi} \|\tilde{y}(0) - \tilde{z}(0)\| \\ &\quad + \int_0^t e^{\int_\xi^t \tilde{C}(\eta)d\eta} \|\tilde{f}(\xi, \tilde{z}(\xi)) - \tilde{g}(\xi, \tilde{z}(\xi))\|\, d\xi \\ &= e^{\int_0^t C(x_0 - \xi)d\xi} \|y(x_0 - 0) - z(x_0 - 0)\| \\ &\quad + \int_0^t e^{\int_\xi^t C(x_0 - \eta)d\eta} \|\tilde{f}(\xi, \tilde{z}(\xi)) - \tilde{g}(\xi, \tilde{z}(\xi))\|\, d\xi.\end{aligned}$$

For the first term, we do a change of variables; ξ from 0 to t, put $s = x_0 - \xi$; $d\xi = -ds$. $0 \le \xi \le t$; s from $x_0 - 0$ to $x_0 - t = x_0 - (x_0 - x) = x$. We conclude that the first term is equal to

$$e^{-\int_{x_0}^{x} C(s)ds} \|y(x_0) - z(x_0)\| = e^{\int_{x}^{x_0} C(s)ds} \|y(x_0) - z(x_0)\|.$$

For the second term, similar changes of variables can be done, resulting in Corollary A.3. □

APPENDIX B: MIMICKING COUNTERFACTUAL OUTCOMES

In this appendix we present conditions under which $X(t)$ mimics $Y^{(t)}$ in the sense that it has the same distribution as $Y^{(t)}$ given $\overline{Z}_t$. This result is used heavily in this article. For the proofs, which are lengthy and use discretization, we refer to Lok (2001, 2004). Section B.2 deals with survival outcomes, Section B.1 with other outcomes. Survival outcomes require a different set of assumptions, as will become clear below. The conditions here are somewhat more restrictive than the ones in Lok (2001, 2004), but they are simpler.

B.1. Mimicking counterfactual nonsurvival outcomes. This section contains a sufficient set of regularity conditions to have existence and uniqueness of a solution $X(t)$ to (6), $X'(t) = D(X(t), t; \overline{Z}_t)$ with final condition $X(\tau) = Y$, the observed outcome (see Figure 3). Furthermore, together with Assumption 6.1 (consistency), they imply that $X(t)$ has the same distribution as $Y^{(t)}$ given $\overline{Z}_t$.

The regularity conditions below should be read as the following: there exist conditional distribution functions $F_{Y^{(t+h)}|\overline{Z}_t}$ such that all these assumptions are satisfied. They can be relaxed to h in a neighborhood of 0, if this neighborhood does not depend on $\overline{Z}$. We only consider $h \ge 0$, so the derivative with respect to h at $h = 0$ is always the right-hand derivative.

ASSUMPTION B.1 (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 $f_{Y^{(t)}|\overline{Z}_t}(y) \ge \varepsilon$ 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 $[y_1, y_2] \times [t_1, t_2) \times \mathbb{R}$ 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.

The support conditions may be restrictive for certain applications. Nevertheless, most real-life situations can be approximated this way, since y_1 and y_2 are unrestricted and $\varepsilon > 0$ is unrestricted, too. Although the support conditions may well be stronger than necessary, they simplify the analysis considerably and, for that reason, they are adopted here. The smoothness conditions allow for nonsmoothness 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 patient and thus to nonsmoothness of the functions concerned.

THEOREM B.2 (Mimicking counterfactual outcomes). *Suppose that regularity Condition B.1 is satisfied. Then $D(y, t; \overline{Z}_t)$ exists. Furthermore, for every $\omega \in \Omega$, there exists exactly one continuous solution $X(t)$ to $X'(t) = D(X(t), t; \overline{Z}_t)$ with final condition $X(\tau) = Y$. If also Assumption 6.1 (consistency) is satisfied and there are no more than finitely many times t for which the probability that the covariate- and treatment process jumps at t is greater than 0, then this $X(t)$ has the same distribution as $Y^{(t)}$ given $\overline{Z}_t$ for all $t \in [0, \tau]$.*

For a proof we refer to Lok (2001, 2004).

B.2. Mimicking counterfactual survival outcomes. This section contains a sufficient set of regularity conditions to have existence and uniqueness of a solution $X(t)$ to equation (6), $X'(t) = D(X(t), t; \overline{Z}_t)$ with final condition $X(\tau) = Y$, the observed outcome (see Figure 3). Furthermore, together with Assumption 6.1 (consistency) and Assumptions B.3 and B.4 below, they imply that $X(t)$ has the same distribution as $Y^{(t)}$ given $\overline{Z}_t$. The conditions here are natural conditions if the outcome of interest Y is a survival time.

As compared to Section B.1, we make two extra assumptions. The first is a consistency assumption, stating that stopping treatment after death does not change the survival time. The second assumption states that there is no instantaneous effect of treatment at the time the patient died [notice that the difference between $Y^{(Y)}$, the outcome with treatment stopped at the survival time Y, and Y is in treatment *at* Y].

ASSUMPTION B.3 (Consistency). $Y^{(t)} = Y$ on $\{\omega : Y < t\} \cup \{\omega : Y^{(t)} < t\}$.

ASSUMPTION B.4 (No instantaneous effect of treatment at the time the patient died). $Y^{(t)} = Y$ on $\{\omega : Y = t\} \cup \{\omega : Y^{(t)} = t\}$.

Under these assumptions, treatment in the future does not cause or prevent death at present or before:

LEMMA B.5. *Under Assumptions* B.3 *and* B.4:

(a) *For all* $h \geq 0$: $Y^{(t+h)} = Y$ *on* $\{\omega : Y \leq t\} \cup \bigcup_{h \geq 0} \{\omega : Y^{(t+h)} \leq t\}$.

(b) *For all* (y, t, h) *with* $y \leq t + h$ *and* $h \geq 0$: $\{\omega : Y^{(t+h)} \leq y\} = \{\omega : Y \leq y\}$.

For a proof we refer to Lok (2001).

If the outcome is survival, the support condition in Assumption B.1, saying that all $F_{Y^{(t+h)}|\overline{Z}_t}$ have the same bounded support $[y_1, y_2]$, will not hold. The reason for this is as follows. $\overline{Z}_t$ includes the covariate-measurements and treatment until time t. If covariates and treatment were measured at time t, it cannot be avoided to include in $\overline{Z}_t$ whether or not a patient was alive at time t. Given that a patient is dead at time t and given his or her survival time, the distribution of this survival time cannot have the fixed support $[y_1, y_2]$, which is independent of t. Also, given that a patient is alive at time t, this is hardly ever the case; one often expects that t is the left limit of the support. Thus, in case the outcome is survival, the support condition for Theorem B.2 has to be slightly changed.

ASSUMPTION B.6 (*Support*). There exists a finite number $y_2 \geq \tau$ such that:

(a) For every $\omega \in \Omega$ and t with $Y > t$, all $F_{Y^{(t+h)}|\overline{Z}_t}$ for $h \geq 0$ have support $[t, y_2]$.

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

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

Next we look at the differentiability conditions in Assumption B.1. It does not seem reasonable to assume that $F_{Y^{(t+h)}|\overline{Z}_t}(y)$ is continuously differentiable with respect to h and y on $(h, y) \in [0, \infty) \times [t, y_2]$ since, for $y \leq t + h$, $F_{Y^{(t+h)}|\overline{Z}_t}(y) = F_{Y|\overline{Z}_t}(y)$ [Lemma B.5(b)]. Therefore, the derivative of $F_{Y^{(t+h)}|\overline{Z}_t}(y)$ with respect to h is likely not to exist at $y = t + h$ (and is equal to zero for $y < t + h$). Also, the derivative of $F_{Y^{(t+h)}|\overline{Z}_t}(y)$ with respect to y may not exist at $y = t + h$, because of the different treatment before and after $t + h$. For survival outcomes, we replace the smoothness conditions of Assumption B.1 by the following:

ASSUMPTION B.7 (*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 $F_{Y^{(t+h)}|\overline{Z}_t}(y)$ with respect to y and h are bounded by constants C_1 and C_2, respectively, for $y \in [t + h, y_2]$.

(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, for $y \in [t+h, y_2]$.

The smoothness condition above concentrates on $y \geq t+h$. For $y \in [t, t+h)$ we can choose $F_{Y^{(t+h)}|\overline{Z}_t}(y) = F_{Y|\overline{Z}_t}(y)$ because of Lemma B.5(b). Because of Assumption 6.1 (consistency), $F_{Y|\overline{Z}_t}$ has the same support as $F_{Y^{(\tau)}|\overline{Z}_t}$, so $F_{Y|\overline{Z}_t}$ has support $[t, y_2]$ if $Y > t$ [Assumption B.6(a)]. Assume the following

ASSUMPTION B.8 (*Smoothness*). 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 B.9 (Mimicking counterfactual survival outcomes). *Suppose that regularity Conditions B.6, B.7 and B.8 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 $X'(t) = D(X(t), t; \overline{Z}_t)$ with final condition $X(\tau) = Y$. If also Assumptions 6.1, B.3 and B.4 (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 $t \in [0, \tau]$.*

For a proof we refer to Lok (2001, 2004).

APPENDIX C: MEASURABILITY ISSUES

In most of this article we assume that the function which maps $(X(t), \overline{Z}_{t-})$ to $X(t_0)$, with $t_0 < t$, is a measurable function on $\mathbb{R} \times \overline{\mathcal{Z}}_{t-}$, with the projection σ-algebra on $\overline{\mathcal{Z}}_{t-}$ (see Section 2). Moreover, we sometimes assume that the function which maps $(X(t_0), \overline{Z}_{t-})$ to $X(t)$, with $t_0 < t$, is a measurable function on $\mathbb{R} \times \overline{\mathcal{Z}}_{t-}$. In this appendix we give sufficient conditions for this. If these two functions are measurable, $\sigma(\overline{Z}_t, X(t))$ is a filtration, and, moreover, $\sigma(\overline{Z}_t, X(t))$ is the same as $\sigma(\overline{Z}_t, X(0))$ [see equation (15) in Section 9].

LEMMA C.1. *Suppose that D satisfies regularity Assumption 9.1 and that, for each $\omega \in \Omega$, Z jumps at most finitely many times. Then the function which maps $(X(t), \overline{Z}_{t-})$ to $X(t_0)$, with $t_0 < t$, is a measurable function from $\mathbb{R} \times \overline{\mathcal{Z}}_{t-}$ to $\mathbb{R}$. Also, the function which maps $(X(t_0), \overline{Z}_{t-})$ to $X(t)$, with $t_0 < t$, is a measurable function from $\mathbb{R} \times \overline{\mathcal{Z}}_{t-}$ to $\mathbb{R}$.*

For the proof of this result, which is quite technical since many results on differential equations are nonconstructive, we refer to Lok (2001). The proof uses the idea behind Euler's forward method to approximate the solution to the differential equation.

APPENDIX D: A CONVERGENCE RESULT

The following lemma is a worked-out case of theory from Van der Vaart (1998), Chapter 19.

LEMMA D.1. *Let $X_1, X_2, \ldots$ be i.i.d. random variables with values in a measurable space $\mathcal{X}$. Let $\{f_\theta : \theta \in \Theta\}$ be a collection of measurable functions from $\mathcal{X}$ to $\mathbb{R}^k$ indexed by a subset $\Theta \subset \mathbb{R}^d$ which contains an open neighborhood Θ_0 of θ_0. Suppose that $\theta \to f_\theta(x)$ is continuous on Θ_0 for every $x \in \mathcal{X}$. Suppose also that there exists a measurable function F on $\mathcal{X}$ such that $\|f_\theta\| \le F$ for every $\theta \in \Theta_0$ and such that $EF(X_1)$ exists. Then if $\hat{\theta}$ converges in probability to θ_0,*

$$P_n f_{\hat{\theta}} \xrightarrow{P} E f_{\theta_0}(X_1),$$

where P_n indicates the empirical distribution of $X_1, X_2, \ldots, X_n$.

PROOF. Notice that

$$\|P_n f_{\hat{\theta}} - E f_{\theta_0}(X_1)\| \le \|P_n f_{\hat{\theta}} - E f_{\hat{\theta}}(X_1)\| + \|E f_{\hat{\theta}}(X_1) - E f_{\theta_0}(X_1)\|.$$

We show that both terms converge to zero in probability. Choose $\Theta_1 \subset \Theta_0$ compact and such that it contains an open neighborhood of θ_0. Example 19.8 from Van der Vaart (1998) implies that, under the conditions above,

$$\sup_{\theta \in \Theta_1} \|P_n f_\theta - E f_\theta(X_1)\| \to 0 \qquad \text{a.s.}$$

Since $\hat{\theta} \xrightarrow{P} \theta_0$ and Θ_1 contain an open neighborhood of θ_0, this implies that the first term converges in probability to zero. For the second term, notice that, on Θ_0, $\theta \to f_\theta(x)$ is continuous in θ and that each of the components of f_θ is bounded by the integrable function F, so that Lebesgue's dominated convergence theorem implies that $E f_\theta(X_1)$ is continuous in θ on Θ_0. Thus, since $\hat{\theta} \xrightarrow{P} \theta_0$, also the second term converges in probability to zero. □

Acknowledgments. An earlier version of this article was a chapter of my Ph.D. dissertation at the Free University of Amsterdam. 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. I thank both referees for helpful suggestions. Besides that, I thank the Netherlands Organization for Scientific Research (NWO) for a scholarship.

REFERENCES

ANDERSEN, P. K., BORGAN, Ø., GILL, R. D. and KEIDING, N. (1993). *Statistical Models Based on Counting Processes*. Springer, New York. MR1198884

BILLINGSLEY, P. (1986). *Probability and Measure*. Wiley, New York. MR0830424

COLLETT, D. (1994). *Modelling Survival Data in Medical Research*. Chapman and Hall, London.

COX, D. R. and HINKLEY, D. V. (1974). *Theoretical Statistics*. Chapman and Hall, London. MR0370837

COX, D. R. and OAKES, D. (1984). *Analysis of Survival Data*. Chapman and Hall, London. MR0751780

DUISTERMAAT, J. J. and ECKHAUS, W. (1995). *Analyse van Gewone Differentiaalvergelijkingen*. Epsilon, Utrecht.

GILL, R. D. and ROBINS, J. M. (2001). Causal inference for complex longitudinal data: The continuous case. *Ann. Statist.* **29** 1785–1811. MR1891746

HERNÁN, M. A., COLE, S. R., MARGOLICK, J., COHEN, M. and ROBINS J. M. (2005). Structural accelerated failure time models for survival analysis in studies with time-varying treatments. *Pharmacoepidemiology and Drug Safety* **14** 477–491.

KEIDING, N. (1999). Event history analysis and inference from observational epidemiology. *Statistics in Medicine* **18** 2353–2363.

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.

LOK, J. J. (2001). Statistical modelling of causal effects in time. Ph.D. thesis, Dept. Mathematics, Free Univ. Amsterdam. Available at http://www.math.vu.nl/research/theses/pdf/lok.pdf.

LOK, J. J. (2004). Mimicking counterfactual outcomes for estimation of causal effects. Available at http://arXiv.org/abs/math.ST/0409045.

LOK, J. J. (2007). Structural nested models and standard software: A mathematical foundation through partial likelihood. *Scand. J. Statist.* **34** 186–206. MR2325250

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. *Statist. Neerlandica* **58** 271–295. MR2157006

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. *Statistics in Medicine* **12** 1605–1628.

PEARL, J. (2000). *Causality. Models*, *Reasoning*, *and Inference*. Cambridge Univ. Press. MR1744773

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, DC.

ROBINS, J. M. (1992). Estimation of the time-dependent accelerated failure time model in the presence of confounding factors. *Biometrika* **79** 321–334. MR1185134

ROBINS, J. M. (1993). Analytic methods for HIV treatment and cofactor effects. In *Methodological Issues of AIDS Behavioral Research* (D. G. Ostrow and R. Kessler, eds.) 213–287. Plenum Press, New York.

ROBINS, J. M. (1997). Causal inference from complex longitudinal data. *Latent Variable Modeling and Applications to Causality*. *Lecture Notes in Statist.* **120** 69–117. Springer, New York. MR1601279

ROBINS, J. M. (1998a). Correcting for non-compliance in equivalence trials. *Statistics in Medicine* **17** 269–302.

ROBINS, J. M. (1998b). Structural nested failure time models. In *Survival Analysis* (P. Armitage and T. Colton, eds.). *Encyclopedia of Biostatistics* **6** 4372–4389. Wiley, Chichester.

ROBINS, J. M. (2000). Robust estimation in sequentially ignorable missing data and causal inference models. In *Proceedings of the American Statistical Association Section on Bayesian Statistical Science* **1999** 6–10.

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.

ROBINS, J. M., ROTNITZKY, A. and SCHARFSTEIN, D. O. (2000). Sensitivity analysis for selection bias and unmeasured confounding in missing data and causal inference models. In *Statistical Models in Epidemiology*, *The Environment and Clinical Trials* (M. E. Halloran and D. Berry, eds.) 1–94. Springer, New York. MR1731681

ROBINS, J. M. and WASSERMAN, L. (1997). Estimation of effects of sequential treatments by reparameterizing directed acyclic graphs. In *Proceedings of the Thirteenth Conference on Uncertainty in Artificial Intelligence* (D. Geiger and P. Shenoy, eds.) 409–420. Morgan Kaufmann, San Francisco, CA.

ROGERS, L. C. G. and WILLIAMS, D. (1994). *Diffusions*, *Markov Processes*, *and Martingales*. Wiley, New York. MR1331599

ROSENBAUM, P. R. and RUBIN, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika* **70** 41–55. MR0742974

VAN DER VAART, A. W. (1998). *Asymptotic Statistics*. Cambridge Univ. Press. MR1652247

VAN DER VAART, A. W. and WELLNER, J. A. (1996). *Weak Convergence and Empirical Processes*. Springer, New York. MR1385671

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 Study. *American J. Epidemiology* **148** 390–401.

DEPARTMENT OF BIOSTATISTICS
HARVARD SCHOOL OF PUBLIC HEALTH
BUILDING II, ROOM 409
655 HUNTINGTON AVENUE
BOSTON, MASSACHUSETTS 02115
USA
E-MAIL: jlok@hsph.harvard.edu