Predicting time-to-event outcomes under different intervention strategies: methods and applications

Prosepe, I.

Chapter 6

Causal multistate models to evaluate treatment delay

Ilaria Prosepe, Saskia le Cessie, Hein Putter, Nan van Geloven

Abstract

Multistate models allow for the study of scenarios where individuals experience different events over time. While effective for descriptive and predictive purposes, multistate models are not typically used for causal inference. We propose an estimator that combines a multistate model with g-computation to estimate the causal effect of treatment delay strategies. In particular, we estimate the impact of strategies such as awaiting natural recovery for 3 months, on the marginal probability of recovery. We use an illness-death model, where illness and death represent respectively treatment and recovery. We formulate the causal assumptions needed for identification and the modeling assumptions needed to estimate the quantities of interest. In a simulation study, we present scenarios where the proposed method can make more efficient use of data compared to an alternative approach using cloning-censoring-reweighting. We then showcase the proposed methodology on real data by estimating the effect of treatment delay on a cohort of 1896 couples with unexplained subfertility who seek intrauterine insemination.

Statistics in Medicine, 44(7). DOI: 10.1002/sim.70061

6.1 Introduction

Multistate models have gathered considerable attention in medical statistics, offering a pivotal extension of survival analysis methodology to settings where individuals may experience multiple different events over time [1–3]. Multistate models are primarily used for descriptive and predictive purposes, e.g. describing the likelihood of a specific event occurring in the presence of intermediate or competing events or gaining insight into the association between prognostic factors and different transition probabilities. Transitions in the multistate model represent those observed in the data. Thus, when used on observational data, multistate models are not targeted towards answering causal questions that aim to estimate outcomes under "what if" scenarios resulting from potential interventions that change certain transitions in the model.

The value of estimating transition rates under hypothetical changes in event distributions was heavily debated in survival analysis literature, as identification requires additional assumptions [4–9]. In particular, survival in a hypothetical world where one competing cause of death is removed is identifiable under the assumptions that a realization of the latent failure times is unchanged by cause removal and that latent failure times are independent of one another [4, 5, 8]. Formulating these assumptions necessitates a comprehensive understanding of the mechanism under study and demands careful consideration of the setting in which the data was collected. This led many researchers to advocate for "sticking to this world" and methodological extensions of the multistate framework predominantly followed this advice by restricting to *observable* quantities rather than latent failure times [4, 5, 8, 10, 11].

It is nowadays increasingly recognized that when the research question is intrinsically causal, the analysis should reflect this, with careful description and assessment of the assumptions [12]. Some recent works applied a causal approach in a multistate setting. Gran et al. [13] investigated the effect of hypothetical interventions on return-to-work for a cohort of work rehabilitation participants, showing how multistate models can be employed to address causal questions: by including treatment/exposure as one of the states, by modeling treatment/exposure as a covariate that has an effect on some of the transitions, or by reweighting the population to estimate the transition intensities in a different population. Valeri et al. [14] recently examined how inequities in access to healthcare contribute to racial disparities in the survival of cancer patients, by employing a multistate model for the hypothetical scenario where a black person would have the same access to healthcare as a white person. Young et al. [15] formulated a causal framework for classical statistical estimands in failure time settings with competing events for contexts where treatment is assigned at baseline. Erdmann et al. [16] used multistate models to estimate treatment effects in the hypothetical scenario of a clinical trial where the treatment was never interrupted due to patients no longer being allowed to receive the investigational drug.

In this paper, we propose the use of an illness-death model, a specific type of multistate model, to estimate the causal effect of treatment delay from observational data in the presence of baseline confounders. In the proposed illness-death model, treatment and recovery represent respectively the analogue to illness and death. We provide estimands for when hypothetical modifications to the transition from the starting state to the treatment state are of interest. Investigating the impact of treatment delay has clear medical relevance, as it may help to prevent potentially expensive and invasive treatments for patients who stand a reasonable chance of

recovering without intervention. However, the optimal delay of treatment initiation, often referred to as a "wait-and-see" or "expectant management" periods, remains largely unknown. Our multistate formulation shares similarities to that of Valeri et al.[14]. However, the estimands of interest and the formulation differ, as they approach their problem from a mediation analysis standpoint [14].

The rest of the paper is set up as follows: in Section 2 we provide a formal definition of the recovery probabilities in the hypothetical scenario of a fixed delay of treatment initiation, outline the assumptions required to identify this quantity from observational data in the presence of baseline confounding, and introduce our proposed estimation approach; in Section 3 we present a simulation study to assess the performance of our proposed method in small-sample scenarios and compare it to the clone-censor-reweighting method [17, 18]; in Section 4 we apply our method to a cohort of couples with unexplained subfertility, studying timing of intrauterine insemination. In Section 5 we discuss our findings.

6.2 Methods

6.2.1 Setting

We consider the situation outlined in Figure 6.1. Initially, all patients are untreated at time 0 (starting state), which may correspond to diagnosis or, more generally, the moment when patients consult a doctor for guidance on whether and when to initiate treatment. At this point, doctor and patient choose their treatment strategy: some patients start treatment right away, while others choose to delay treatment initiation to first see if recovery without treatment occurs. In this context, patients may transition between three states: starting state (state 1), treatment (state 2), and recovery (state 3). Some move directly from starting state to recovery, while others first transition from starting state to treatment. Not all patients necessarily reach the recovery state: some may never leave the starting state, and some may never leave the treatment state. This setting is known as an irreversible *illness-death model* in multistate literature [2]. This paper focuses on studying treatment strategies of the form "if not recovered by a certain time, then initiate treatment." We refer to the waiting time before treatment initiation as "treatment delay".

6.2.2 Notation and estimand

Let s denote the time since entry on the starting state ($s = 0$). We let $\tilde{R}$ be the time from baseline to recovery. Some patients transition directly from starting state to recovery, while others first receive treatment. We let $\tilde{W}$ be the time to first event (so either recovery if the patient recovers untreated or treatment if the patient receives treatment first). Following a *competing risks* notation, we define the indicator $\delta_{\tilde{W}}$ which takes value 1 if the patient is treated and takes value 2 if the patient recovers without treatment. We then let C be time from baseline to censoring. We let $R = \min\{\tilde{R}, C\}$ be the observed recovery time, with status indicator δ_R which takes value 1 if we observe the recovery and 0 if the patient is censored

before recovery. We let $W = \min\{\tilde{W}, C\}$ be the observed first event with indicator δ_W which takes value 1 if the patient is treated, takes value 2 if the patient recovers without treatment and and takes value 0 if the patient is censored before either treatment or recovery. Times R and W depend on X, a vector of baseline covariates that confound the relation between treatment time and recovery.

We let g be a treatment strategy of the form "if not yet recovered by time t_g, then initiate treatment at t_g." The strategy "never assign treatment" corresponds to $t_g = \infty$. We define R^g as the potential recovery time under treatment strategy g. Our estimand of interest is $\text{Prob}(R^g < \ell)$, the marginal, i.e. population averaged, recovery probability by a fixed time point ℓ under treatment strategy g.

6.2.3 Multistate model

We define the (observable) hazard rates, also known as *transition intensities* [2], for the three transitions as

$$\begin{aligned}
\lambda_{12}(s \mid X) &= \lim_{\Delta s \to 0} \frac{\text{Prob}(s \leq W < s + \Delta s,\ \delta_W = 1 \mid W \geq s, X)}{\Delta s} \\
\lambda_{13}(s \mid X) &= \lim_{\Delta s \to 0} \frac{\text{Prob}(s \leq W < s + \Delta s,\ \delta_W = 2 \mid W \geq s, X)}{\Delta s} \\
\lambda_{23,t}(s \mid X) &= \lim_{\Delta s \to 0} \frac{\text{Prob}(s \leq R < s + \Delta s, \delta_R = 1 \mid R \geq s, W = t, \delta_W = 1, X)}{\Delta s} \text{ with } s \geq t.
\end{aligned} \tag{6.1}$$

We define the cumulative hazard function for the $i \longrightarrow j$ transitions as

$$\begin{aligned}
\Lambda_{1j}(s \mid X) &= \int_0^s \lambda_{1j}(u \mid X) \mathrm{d}u \quad \text{for } j \in \{2, 3\} \\
\Lambda_{23,t}(s \mid X) &= \int_t^s \lambda_{23,t}(u \mid X) \mathrm{d}u \quad \text{with } s \geq t
\end{aligned} \tag{6.2}$$

and the functions $S_{12}(s, X)$, $S_{13}(s, X)$ and $S_{23,r}(s \mid X)$ as

$$\begin{aligned}
S_{1j}(s \mid X) &= \exp(-\Lambda_{1j}(s \mid X)) = \prod_{u \leq s} (1 - \lambda_{1j}(u \mid X))\, \mathrm{d}u \quad \text{for } j \in \{2, 3\} \\
S_{23,t}(s \mid X) &= \exp(-\Lambda_{23,t}(s \mid X)) = \prod_{t \leq u \leq s} (1 - \lambda_{23,t}(u \mid X))\, \mathrm{d}u \quad \text{with } s \geq t.
\end{aligned} \tag{6.3}$$

where $\prod_{u \leq s}$ indicates the product taken over the infinitesimally small intervals $[s, s + \Delta s]$. We remark that the functions $S_{1j}(s, X)$ ($j \in \{2, 3\}$), while they can be estimated, can not in general be interpreted as survival distributions. Only when the competing events (and the censoring process) are independent conditionally on X, can these functions be interpreted as the survival distribution in the situation that the competing event does not occur. We refer to literature[2] for further background.

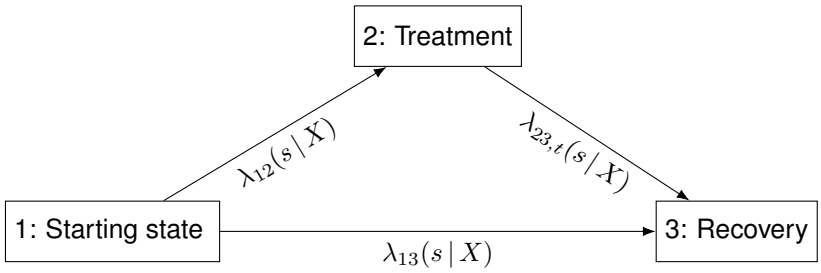

Figure 6.1: The illness-death setting: the three states (starting state, treatment and recovery) are connected by arrows that represent the possible transitions between them. The respective transition intensities between states are depicted.

6.2.4 Identification of causal effects for multistate models

In the following, we demonstrate how our quantity of interest $\text{Prob}(R^g < \ell)$ is identifiable under four identifiability conditions: consistency and, conditional on a set of baseline covariates X, positivity, exchangeability and independent censoring. These conditions, delineated in this section, represent an adapted version of the standard identifiability conditions used in causal inference for time-to-event outcomes [19], to match our setting of interest. Under these assumptions, the following identification formula holds:

$$\text{Prob}(R^g < \ell) = \mathbf{E}_X \begin{cases} 1 - S_{13}(\ell \,|\, X) & \text{for } \ell \leq t_g \\ 1 - S_{13}(t_g \,|\, X) \cdot S_{23,t_g}(\ell \,|\, X) & \text{for } \ell > t_g. \end{cases} \tag{6.4}$$

We will now formulate in detail the four identifiability conditions and prove Equation (6.4) under these conditions.

Let $A(s)$ be the time-dependent treatment indicator that takes the value 0 if the patient is still untreated at time s and 1 otherwise and let $\bar{A}(s)$ be the treatment history up to s: $\bar{A}(s) = \{A(u) : u \in [0, s]\}$. Similarly, let $C(s)$ be the time-dependent censoring indicator that takes the value 0 if the patient is uncensored at time s and 1 otherwise, and let $\bar{C}(s)$ be the censoring history up to s: $\bar{C}(s) = \{C(u) : u \in [0, s]\}$. Under this notation, the identifiability conditions are:

- **Consistency:** $R^g = (\tilde{R} \,|\, A(\tilde{R}) = 0)$ if $\tilde{R} \leq t_g$ and $R^g = (\tilde{R} \,|\, \tilde{W} = t_g, \delta_{\tilde{W}} = 1)$ if $R > t_g$. This means that the counterfactual outcome R^g is equal to the outcome $\tilde{R}$ on those subjects who actually follow the treatment strategy g.
- **Positivity:** $\lim_{\Delta s \longrightarrow 0} \frac{1}{\Delta s} \text{Prob}(t_g \leq W < t_g + \Delta s, \delta_W = 1 \,|\, W \geq t_g, X) > 0$.
  In words, the treatment rate at time t_g given X for subjects who have not yet recovered, have not yet been treated and are not yet censored is larger than 0, for every treatment strategy g of interest and for all observed values of X. This also implies that there are still subjects at risk for transitions $1 \to 2$ and $1 \to 3$ at time t_g.

- **Conditional exchangeability:** for $\Delta s \to 0$,

$$I(s \leq R^g < \Delta s) \perp\!\!\!\perp \bar{A}(s+\Delta s) \,|\, \tilde{R} \geq s, X, \{A(u) = I(u > t_g)\ \forall\ u \in [0,s]\}.$$

  In words, within each level of X, the probability of being treated during the interval $[s, s+\Delta s)$ is independent of the outcome under treatment strategy g in that same interval, for all $s \geq 0$. The conditional exchangeability assumption requires that, within each level of the baseline confounder X, treated and untreated patients are comparable at each treatment time, meaning those who receive treatment are representative of those who do not (and vice versa). This assumption holds when the decision to initiate treatment is made based on baseline characteristics only, as outlined in Section 6.2.1. If the decision to initiate treatment is based on time-varying confounders, this assumption is not met.

- **Independent censoring**: for $\Delta s \to 0$:

  (a) $I(s \leq \tilde{R} < \Delta s) \perp\!\!\!\perp \bar{C}(s+\Delta s) \,|\, \{A(u) = I(u > t_g)\ \forall\ u \in [0,s]\}, R \geq s, X$;
  (b) for $s \leq t_g$, $\bar{A}(s+\Delta s) \perp\!\!\!\perp \bar{C}(s+\Delta s) \,|\, W \geq s, X$.

  In words, $\tilde{R}$ and the observed treatment strategy are independent of the censoring mechanism conditionally on X. Similarly to the conditional exchangeability assumption, the independent censoring assumption requires that, within each level of X, the patients who remain uncensored at each censoring time are representative of those who were censored.

Let us now denote by $\lambda^g(s \,|\, X)$ the counterfactual hazard rate of recovery under treatment strategy g:

$$\lambda^g(s|X) = \lim_{\Delta s \to 0} \frac{\text{Prob}(s \leq R^g < s+\Delta s \,|\, R^g \geq s, X)}{\Delta s}. \tag{6.5}$$

Under this notation

$$\text{Prob}(R^g < \ell) = \mathbf{E}_X \text{Prob}(R^g < \ell \,|\, X) = \mathbf{E}_X \left(1 - \prod_{s \leq \ell} (1 - \lambda^g(s \,|\, X) \mathrm{d}s)\right) \tag{6.6}$$

where the first equality is due to the law of total expectation and the second follows by splitting the interval $[0, s]$ into infinitesimally small sub-intervals. By assuming consistency, positivity and conditional exchangeability, for $s \leq t_g$, we have

$$\begin{aligned}
\lambda^g(s \,|\, X) \;&:=\; \lim_{\Delta s \to 0} \frac{\text{Prob}(s \leq R^g < s+\Delta s \,|\, R^g \geq s, X)}{\Delta s} \\
&\overset{\text{C.E.}}{=} \lim_{\Delta s \to 0} \frac{\text{Prob}(s \leq R^g < s+\Delta s \,|\, R^g \geq s, A(s^-) = 0, X)}{\Delta s} \\
&\overset{\text{Co.}}{=} \lim_{\Delta s \to 0} \frac{\text{Prob}(s \leq R^g < s+\Delta s \,|\, \tilde{W} \geq s, X)}{\Delta s} \\
&\overset{\text{C.E. + Co.}}{=} \lim_{\Delta s \to 0} \frac{\text{Prob}(s \leq \tilde{W} < s+\Delta s, \delta_{\tilde{W}} = 2 \,|\, \tilde{W} \geq s, X)}{\Delta s} \\
&\overset{\text{I.C.}}{=} \lim_{\Delta s \to 0} \frac{\text{Prob}(s \leq \tilde{W} < s+\Delta s, \delta_{\tilde{W}} = 2 \,|\, W \geq s, X)}{\Delta s} \\
&\overset{\text{I.C.+ def. W}}{=} \lim_{\Delta s \to 0} \frac{\text{Prob}(s \leq W < s+\Delta s, \delta_{W} = 2 \,|\, W \geq s, X)}{\Delta s} \;=:\; \lambda_{13}(s \,|\, X).
\end{aligned} \tag{6.7}$$

Conditional exchangeability (C.E.) ensures the first equality as it allows to condition on patients who are still untreated just before time s. Consistency (Co.) ensures the second equality as it allows to switch from the condition $R^g \geq s$ to $R \geq s$, which combined with $A(s) = 0$ yields the condition $W \geq s$. Conditional exchangeability ensures the third equality of Equation (6.7) as it allows restricting to patients who remain untreated during the next short amount of time Δs and, together with consistency, it ensures we can switch from the counterfactual $s \leq R^g < s + \Delta s$ to the factual $s \leq \tilde{W} < s + \Delta s, \delta_{\tilde{W}} = 2$. Points (a) and (b) of the independent censoring (I.C.) assumption ensure that $\tilde{R}$, $A(s)$ and $C(s)$ are mutually independent and, therefore, that we can condition on patients who are still uncensored (and untreated) just before time s (fourth equality of Equation (6.7)). Points (a) and (b) of the independent censoring assumption (I.C.) ensure, together with $W = \min\{\tilde{W}, C\}$, the fifth equality of Equation (6.7), as they allow restricting to patients who remain uncensored (and untreated) during the next short amount of time Δs. Positivity ensures that there are subjects at risk for transition $1 \to 3$ for all $s \leq t_g$, ensuring that $\lambda_{13}(s \,|\, X)$ exists for all $s \leq t_g$. This would not happen if, for example, patients with certain X values all received treatment before time t_g. Similarly, for $s > t_g$:

$$
\begin{aligned}
\lambda^g(s \,|\, X) \;&:=\; \lim_{\Delta s \to 0} \frac{\text{Prob}(s \leq R^g < s + \Delta s \,|\, R^g \geq s, X)}{\Delta s} \\
&\overset{\text{C.E.}}{=} \lim_{\Delta s \to 0} \frac{\text{Prob}(s \leq R^g < s + \Delta s \,|\, \tilde{W} = t_g, \delta_{\tilde{W}} = 1, R^g \geq s, X)}{\Delta s} \\
&\overset{\text{Co.}}{=} \lim_{\Delta s \to 0} \frac{\text{Prob}(s \leq \tilde{R} < s + \Delta s \,|\, \tilde{W} = t_g, \delta_{\tilde{W}} = 1, \tilde{R} \geq s, X)}{\Delta s} \\
&\overset{\text{I.C.+ def. W}}{=} \lim_{\Delta s \to 0} \frac{\text{Prob}(s \leq R < s + \Delta s, \delta_R = 1 \,|\, W = t_g, \delta_W = 1, R \geq s, X)}{\Delta s} \;=:\; \lambda_{23,t_g}(s \,|\, X).
\end{aligned}
$$

Conditional exchangeability ensures the first equality as it allows to condition on patients who were treated at time t_g. Consistency ensures the second equality as it allows to switch from the condition $R^g \geq s$ to $R \geq s$. The independent censoring assumption ensures, together with $R = \min\{\tilde{R}, C\}$, the equality of the quantities at lines 3 and 4. Positivity ensures that there are subjects that did transition from state 1 to state 2 at time t_g, ensuring that $\lambda_{23,t_g}(s \,|\, X)$ exists.

This means that, when the identifiability conditions hold, the transition intensities $\lambda_{13}(s \,|\, X)$ and $\lambda_{23,t}(s \,|\, X)$ remain unchanged after the modification of transition $1 \to 2$. It follows that we can replace the quantity $\lambda^g(s \,|\, X)$ in Equation (6.6) by $\lambda_{13}(s \,|\, X)$ for $s \leq t_g$ and by $\lambda_{23,t_g}(s \,|\, X)$ for $s > t_g$, yielding Equation (6.4).

6.2.5 Estimation

We propose a semi-parametric g-computation which relies on the correct specification of the outcome model, i.e. of transition hazards $\lambda_{13}(s \,|\, X)$ and $\lambda_{23,t}(s \,|\, X)$. We model $\lambda_{13}(s \,|\, X)$ and $\lambda_{23,t}(s \,|\, X)$ using a Cox-type model with a clock-reset at time of treatment initiation:

$$
\begin{aligned}
\lambda_{13}(s \,|\, X) &= \lambda_{13,0}(s) \exp\{f_{13}(X; \beta_{13})\} \\
\lambda_{23,t}(s \,|\, X) &= \lambda_{23,t,0}(s - t) \exp\{f_{23}(X, t; \beta_{23}, \gamma)\} \quad \text{with } s \geq t,
\end{aligned}
\tag{6.8}
$$

where $\lambda_{13,0}(s)$ and $\lambda_{23,t,0}(s)$ represent baseline hazards for $\lambda_{13}(s \mid X)$ and $\lambda_{23,t}(s \mid X)$ respectively; β_{13} is the (possibly time-dependent) parameter vector for the effect of X on recovery without treatment; β_{23} is the (possibly time-dependent) parameter vector for the effect of X on recovery after treatment; γ is the parameter vector for the effect of treatment delay t on recovery after treatment; and $f_{13}(X;\beta_{13})$ and $f_{23}(X,t;\beta_{23},\gamma)$ are functions representing the (possibly time-dependent) multiplication factor on the log scale for the hazards of transitions $1 \to 3$ and $2 \to 3$ respectively for an individual with covariate values X. With the proposed semi-parametric multistate method, we are less concerned about random violations of the positivity assumption. Unlike non-parametric estimation of (6.4), this method allows us to borrow information across treatment strategies and levels of X, so it is not necessary to observe each treatment strategy g for every level of X. Resetting the clock is helpful for treatment strategies where t_g is close to 0. As all individuals start untreated, data may be insufficient for a correct estimation of $\lambda_{23,t_g}(s \mid X)$ with a clock-forward approach. The flexibility in borrowing information comes at the expense of further assumptions, i.e. the correct specification of $\lambda_{13}(s \mid X)$ and $\lambda_{23,t}(s \mid X)$. A simple choice for these hazards would be using Cox proportional hazards models: $\lambda_{13}(s \mid X) = \lambda_{13,0}(s)\exp\{\beta_{13}X\}$ and $\lambda_{23,t}(s \mid X) = \lambda_{23,t,0}(s-t)\exp\{\beta_{23}X+\gamma t\}$, where the parameter vectors β_{13}, β_{23} and γ do not vary over time. This version of the hazards can be used if the assumptions of linearity and proportional hazards hold. Under this simple choice, the following estimator of (6.4) can be used:

$$\widehat{\text{Prob}}(R^g < \ell) = \mathbf{E}_X \begin{cases} 1 - \exp(-\hat{\Lambda}_{13,0}(\ell)\exp\{\hat{\beta}_{13}X\}) & \text{for } \ell \le t_g \\ 1 - \exp(-\hat{\Lambda}_{13,0}(t_g)\exp\{\hat{\beta}_{13}X\} - \hat{\Lambda}_{23,t,0}(\ell - t_g)\exp\{\hat{\beta}_{23}X + \hat{\gamma}t_g\}) & \text{for } \ell > t_g, \end{cases} \tag{6.9}$$

where the cumulative baseline hazards $\Lambda_{13,0}(s) = \int_0^s \lambda_{13,0}(u)\mathrm{d}u$ and $\Lambda_{23,t,0}(s-t) = \int_t^s \lambda_{23,t,0}(u)\mathrm{d}u$ can be estimated by means of the Breslow estimator and $\mathbf{E}_X$ indicates that we average over the empirical distribution of X in the sample. Interactions can be included in the covariate set X if needed for correct model specification. If the linearity or the proportional hazard assumptions fail for a covariate, an alternative functional form of that covariate or an interaction of that covariate with an appropriate function of time could be employed. We refer to literature [20] for further details on model selection.

6.3 Simulation

We report our simulation set-up and results according to the ADEMP (Aims, Data generating mechanisms, Estimands, Methods, Performance measures) structure [21].

6.3.1 Simulation set-up

Aim

The aim of this simulation is to evaluate the small-sample performance of our proposed method and, secondly, to compare accuracy and efficiency to clone-censor-reweighting, an existing method, based on inverse probability weighting, that could alternatively be used for the estimation of the effect of treatment delay from observational data [17, 18, 22].

Data generating mechanism

We generated four different scenarios. Scenario 1 is the base scenario. Each subsequent scenario differs from the base scenario by introducing one single modification in the data generating mechanism. In all presented scenarios, we generated the data in such a way that the assumptions of consistency, conditional exchangeability and positivity, conditional on the covariate X, hold. Parameter choices are loosely based on the data application presented in Section 6.4. To allow for clearer interpretation of the results, each scenario presented here assumes constant baseline hazards, modeled with an exponential distribution. For comparison, results from the same four scenarios using time-varying baseline hazards from a Weibull distribution are provided in Appendix A.

Scenario 1 - Base: We generated data representing $N = 2500$ patients. We generated one continuous baseline covariate $X \sim \mathcal{N}(0, 1)$, which influences both time-to-treatment and time-to-recovery. We then generated:

- a latent time of recovery without treatment V with hazard $0.4 \cdot \exp(-0.25X)$;
- a latent time of treatment T, drawn from a discrete distribution with $P(T = 0) = \exp(-0.05 \cdot \exp(0.25X))$ and discrete hazard $P(T = s \mid T \geq s) = 0.1 \cdot \exp(0.25X)$ at times $s = \{0.25, 0.5, 0.75, 1\}$; if no treatment time T is drawn, we assume the patient remains untreated until the end of follow-up;
- a latent post-treatment time of recovery U with hazard $0.8 \cdot \exp(-0.15X - 0.25T)$, so that the hazard of recovery decreases if treatment is started later;
- a latent censoring time C with hazard $0.2 \cdot \exp(0.1X)$.

The choice to generate discrete treatment times, and not continuous, for our base scenario was motivated by wanting to ensure practical positivity for all treatment strategies that we wished to study, which is especially needed for clone-censor-reweight approach. The observed final data set was made up of the following covariates: (i) the covariate X; (ii) time of first event $W = \min\{T, V\}$ with treatment status indicator $\delta_T = I(W = T)$; (iii) time of recovery $R = T + U$ if $T < V$ and $R = V$ if $T \geq V$.

The remaining scenarios challenge different assumptions regarding the correct model specification. We modify the data generating mechanism for transition $2 \rightarrow 3$ to challenge our proposed multistate approach, which requires the correct specification of the outcome model, and we modify the transition $1 \rightarrow 2$ to challenge the clone-censor-reweight method, which

requires the correct specification of the time-to-treatment model. The generating mechanism for transition $1 \to 3$ remains unchanged across all scenarios, as modifying transition $2 \to 3$ is enough to challenge models that require the correct specification of the outcome model.

Scenario 2 - Continuous treatment times: Time of treatment T was drawn from a continuous distribution with exponential hazard function $0.4 \cdot \exp(0.25X)$.

Scenario 3 - Non-proportional effect of T in transition $2 \to 3$: The effect of T on the log hazard for U was modelled by separate baseline hazards, one for each discrete treatment time T. The new hazard function of transition $2 \to 3$ followed a Weibull distribution, with the shape parameter that depends on T. This makes the proportional hazard assumption fail for transition $2 \to 3$ with respect to the variable T. We chose the shape parameter $\alpha_T = 0.75 + (0.5 \cdot T)/1.5$, yielding the hazard $0.8 \cdot \alpha_T (s - T)^{\alpha_T - 1} \cdot \exp(-0.15X)$ at time s. This choice regarding the shape parameter α_T was made to obtain recovery rates that are similar to the base scenario, to make it easier to compare results across scenarios.

Scenario 4 - Non-proportional effect of X in transition $1 \to 2$: Instead of a constant effect ($\beta_{12} = 0.25$) of X on the discrete hazard of transition $1 \to 2$, as in the base scenario, we used a time-dependent effect $\beta_{12}(s) = 6(s - 0.5)^2 - 1$, which varies quadratically over time. This makes the proportional hazard assumption fail for transition $1 \to 2$ with respect to the covariate X.

Estimand

The estimand of interest was Prob$(R^g < \ell)$, for $0 < \ell \leq 1.5$, with g a given treatment initiation strategy. In this simulation we compared the following strategies: initiating treatment right away ($t_g = 0$), at $t_g = 0.25$, at $t_g = 0.5$, at $t_g = 0.75$ and not initiating treatment before time 1.5 (we will refer to this strategy as the "Never" strategy).

Methods

We compared the proposed multistate method to the clone-censor-reweight method [17, 18]. To gain deeper insight from this comparison, we considered two variants of each method: one variant assumes that treatment delay is a continuous variable and has a linear effect on the log hazard of the outcome model, while the other categorizes treatment delay and stratifies hazards by treatment delay. All methods rely on the causal identifiability conditions presented in Section 6.2.4.

Multistate continuous: We fitted the multistate model using Cox proportional hazards models, with hazards $\lambda_{13}(s \mid X)$ and $\lambda_{23,t}(s \mid X)$ as specified in Equation (6.8). Treatment delay was included as a continuous variable in the model for the log hazard for the transition from treatment to recovery, assuming linearity. Recovery probabilities under different treatment strategies were then estimated as presented in Equation (6.9). For this method we need correct specification of $\lambda_{13}(s \mid X)$ and $\lambda_{23,t}(s \mid X)$.

Multistate categorical: We modeled $\lambda_{23,t}(s \mid X) = \lambda_{23,t,0}(s - t) \exp\{\beta_{23} X\}$ stratifying the baseline hazard $\lambda_{23,t,0}(s - t)$ by treatment delay t. In scenario 2 where time to treatment is

continuous, treatment delay t was categorized into t_{cat}, with levels $[0, 0.125]$, $[0.125, 0.375]$, $[0.375, 0.625]$, $[0.625, 0.875]$, $[0.875, 1.125]$, $[1.125, 1.5]$. Patients who received treatment after 1.5 (time horizon) were censored at time of treatment to avoid creating baseline hazards based on too few or too incomparable individuals (which are unnecessary for estimating our target estimand). The choice of intervals was aimed at obtaining similar results to scenario 1. This method also relies on the correct specification of $\lambda_{13}(s \mid X)$ and $\lambda_{23,t}(s \mid X)$.

Clone-censor-reweight categorical: We performed: (i) cloning: each individual was assigned to one or more treatment strategies at time zero by creating clones, one for each treatment strategy that was compatible with their observed data at time zero; (ii) censoring: clones were artificially censored when they deviated from their assigned strategy; (iii) reweighting: inverse probability weighting was used to address the dependent censoring we introduced in the data. For the estimation of the weights, a time-to-treatment Cox model was used, assuming a linear and proportional effect of X. In the simulation runs where infinite weights were produced, weights where trimmed to the 97.5 percentile. After cloning, censoring and re-weighting, we estimated the recovery probabilities by means of a (reweighted) Kaplan-Meier estimator. In scenario 2, where time to treatment was continuous, we introduced grace windows spanning 0.125 before and after the target treatment delay, effectively changing the estimand slightly. For example, treatment strategy "start at 0" became "start before time 0.125" and treatment strategy "start at 0.5" became "start between 0.375 and 0.625." This method relies on the correct specification of the time-to-treatment model.

Clone-censor-reweight continuous: After cloning, censoring and reweighting the observational data as described in the previous method, the recovery probabilities were estimated via a (reweighted) Cox model. In the Cox model, the baseline hazard was stratified by treatment yes/no and treatment delay was included linearly as a single continuous variable for the stratum "treatment = yes". Time was reset after treatment initiation. For scenario 2, grace windows spanning 0.125 before and after the target treatment delay were used for cloning, censoring, and reweighting. In the estimation of the reweighted Cox model, the treatment delay variable took the value associated to the treatment strategy of each clone (e.g. for clones who are consistent with the "start before 0.125" grace period, we used treatment delay equal to 0, for clones consistent with "start between 0.375 and 0.625" strategy we used treatment delay equal to 0.5). Unlike the categorical method, this approach did not change the estimand, as we used the Cox model to estimate the probability of recovery for specific point values of treatment delay. This approach relies on the correct specification of both the time-to-treatment model and the Cox model, and assumes that individuals who start treatment within the grace window can be combined together into a single group.

Performance measures

Performance of the four methods was assessed numerically through bias and root-mean-square error (RMSE) at time horizon (1.5) and visually by comparing the true and estimated recovery probabilities along a fine grid of time horizons between 0 and 1.5. The truth was computed as the numerical integral $\int_{-\infty}^{\infty} \text{Prob}(R^g < \ell \mid X = x) f_X(x) \mathrm{d}x$, where $f_X(x)$ is the underlying probability density function of X used in the data generation. We used the hazard $0.8 \cdot \exp(-0.15X - 0.25t_g)$ as true latent post-treatment hazard of recovery under treatment strategy g.

6.3.2 Software

All analyses were conducted using the statistical software R (version 4.3.1) [23] using the packages `mstate` [24], `survival` [25], `tidyverse` [26] and `matrixStats` [27]. Our simulation code is available at:
`https://github.com/survival-lumc/CausalMultistate`.

Table 6.1: Simulation results. Bias and root mean squared error (RMSE) averaged over 200 simulation runs for the estimates of the cumulative recovery probabilities at time 1.5 under six treatment strategies: initiate immediately, delayed at 0.25, delayed at 0.5, delayed at 0.75, delayed at 1 and never initiate (before 1.5). Four estimation approaches are reported: (i) *Multistate continuous*; (ii) *Multistate categorical*; (iii) *Clone-censor-reweight (CCR) continuous*; (iv) *CCR categorical*. Four scenarios are presented: (1) assumptions needed for all estimation approaches are met; (2) time of treatment is generated continuously; (3) the effect of treatment delay on time-to-recovery is non-proportional; (4) the effect of the covariate X on time-to-treatment is non-proportional. Standard errors (SE) of the simulation can be derived using the relation $\text{RMSE}^2 = \text{SE}^2 + \text{Bias}^2$.

Scenario	**Strategy**	**Multistate continuous**		**Multistate categorical**		**CCR continuous**		**CCR categorical**	
		Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE
1 - *Base: correctly specified models*	0 months	0.00	0.03	-0.01	0.05	0.00	0.03	0.00	0.05
	3 months	0.00	0.02	-0.01	0.03	0.00	0.02	0.00	0.03
	6 months	0.00	0.02	-0.01	0.04	0.00	0.02	0.00	0.03
	9 months	0.00	0.01	-0.01	0.04	0.00	0.02	0.00	0.04
	1 year	0.00	0.02	0.00	0.03	0.00	0.02	0.00	0.03
	Never	0.00	0.01	0.00	0.01	0.00	0.01	0.00	0.01
2 - *Continuous treatment times*	0 months	0.00	0.02	-0.01	0.05	-0.01	0.03	-0.02	0.05
	3 months	0.00	0.02	0.00	0.03	0.00	0.03	-0.01	0.04
	6 months	0.00	0.02	0.00	0.04	0.00	0.02	-0.01	0.04
	9 months	0.00	0.01	0.00	0.03	0.00	0.02	-0.01	0.04
	1 year	0.00	0.01	0.00	0.03	0.00	0.02	-0.01	0.03
	Never	0.00	0.01	0.00	0.01	0.00	0.01	0.00	0.01
3 - *Non-proportional effect of T in transition* $2 \to 3$	0 months	0.01	0.02	-0.01	0.04	-0.01	0.03	0.00	0.05
	3 months	0.00	0.02	-0.01	0.03	-0.01	0.02	0.00	0.03
	6 months	0.00	0.02	0.00	0.04	0.00	0.02	0.00	0.04
	9 months	0.01	0.02	0.00	0.03	0.01	0.02	0.00	0.03
	1 year	0.02	0.03	0.00	0.03	0.04	0.04	0.00	0.03
	Never	0.00	0.01	0.00	0.01	0.00	0.01	0.00	0.01
4 - *Non-proportional effect of X in transition* $1 \to 2$	0 months	0.00	0.02	-0.01	0.05	0.03	0.04	-0.01	0.05
	3 months	0.00	0.02	0.00	0.03	0.01	0.02	0.03	0.04
	6 months	0.00	0.01	0.00	0.03	0.00	0.02	0.02	0.04
	9 months	0.00	0.01	0.00	0.03	-0.02	0.02	-0.01	0.04
	1 year	0.00	0.01	0.00	0.02	-0.02	0.03	-0.03	0.04
	Never	0.00	0.01	0.00	0.01	-0.01	0.01	-0.01	0.01

6.3.3 Simulation Results

We ran our simulation 200 times. Table 6.1 provides numerical assessment of bias and RMSE at time horizon 1.5 for all four scenarios. Figure 6.2 provides a visual representation of the

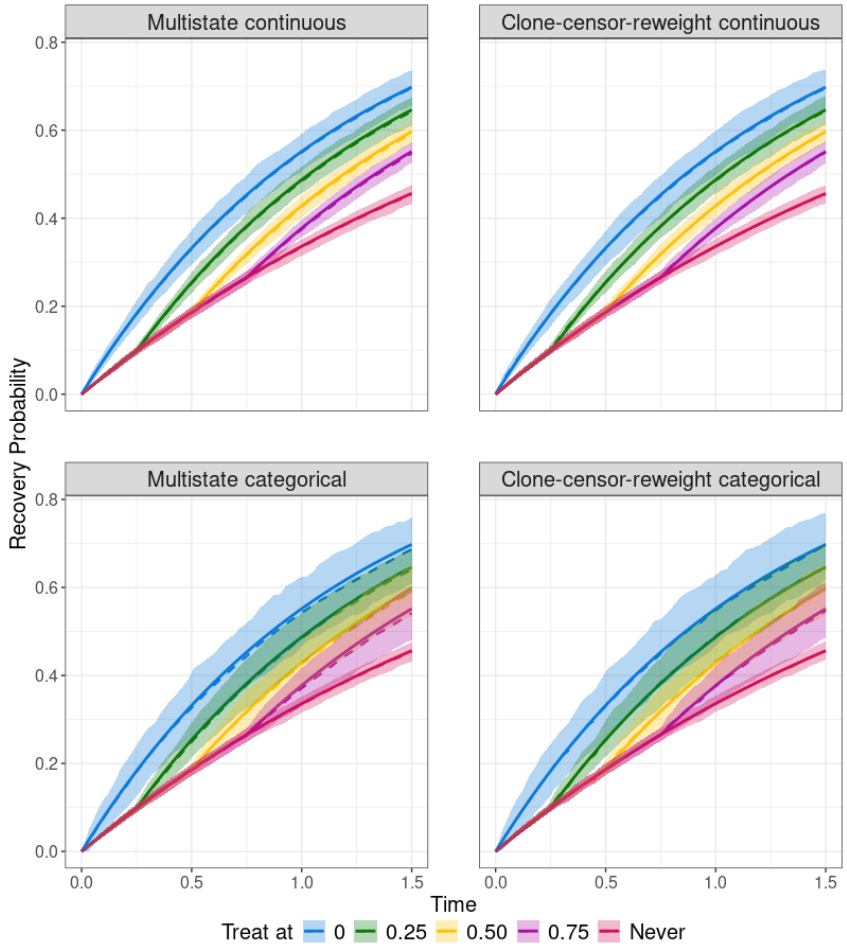

Figure 6.2: Visual representation of the simulation results for scenario 1, where assumptions needed for all estimation approaches are met. Cumulative recovery probabilities until time 1.5 are plotted, under five treatment strategies: initiate immediately, delayed at 0.25, delayed at 0.5, delayed at 0.75, and never initiate. Four estimation approaches are reported: (i) *Multistate continuous*; (ii) *Multistate categorical*; (iii) *Clone-censor-reweight continuous*; (iv) *Clone-censor-reweight categorical*. The solid lines represent the true values, the dotted line represents the mean of the estimates and the shaded area represents the $[5th - 95th]$ percentile of the estimates.

simulation results for the base scenario (scenario 1). Figures for the other scenarios can be found in Appendix A. In the base scenario (scenario 1) all methods performed well. This was expected, as data was generated in a way such that assumptions needed for all four estimation approaches were met. The *multistate continuous* model showed the smallest $[5^{th} - 95^{th}]$ percentile range of the estimates in Figure 6.2 and smallest RMSEs in Table 6.1. The *clone-censor-reweight continuous* ranked second in terms of RMSE.

In scenario 2, where treatment initiation was generated continuously, the *multistate continuous* yielded the smallest bias and RMSE, while the *clone-censor-reweight categorical* method presented bias for strategies where treatment was initiated (see Table 6.1 and Figure 4 in Appendix A). This is consistent with the fact that this method targeted a slightly different estimand. In 47 out of the 200 scenarios examined, the *clone-censor-reweight* method produced infinite weights, which were trimmed. This observation aligns with clone-censor-reweighting being more susceptible to random violations to positivity, which is common when treatment initiation is a continuous variable.

In scenario 3, where treatment delay influenced the log hazard of the recovery chances non-proportionally, the *multistate continuous* and the *clone-censor-reweight continuous* showed bias due to incorrect model specification (see Table 6.1 and Figure 5 in Appendix A). For these methods, the RMSEs at 1.5 increased compared to scenario 1. For the other two methods, the RMSEs remained unchanged from scenario 1.

In the fourth scenario, where the covariate X influenced the log hazard of the recovery chances non-proportionally, the two *clone-censor-reweight* approaches showed increased bias and RMSE compared to the base scenario as the model for the weights were not correctly specified.

6.4 Data application

We applied the proposed method to estimate the probability of getting pregnant within 1.5 years after workup completion under different treatment delay strategies for patients with unexplained subfertility, defined as having tried to conceive naturally for over a year without success [28, 29] despite having parameters of infertility within normal ranges. The interest in determining whether and for how long to delay treatment initiation stems from the fact that while treatment usually increases the pregnancy probabilities [30–32], it also carries potential negative side effects associated with IUI, such as the potential risks of the hormonal therapy accompanying IUI and the financial and psychological burden on the couple.

We used data from a prospective cohort that was recruited across 38 hospitals in The Netherlands between January 2000 and October 2005. A more detailed description of the protocol and of the clinical definitions and setting can be found elsewhere [33, 34]. For the current study, we included couples with unexplained subfertility from seven (out of 38) centers that additionally collected data on intrauterine insemination (IUI). These centers included 1896 couples, with at most 4 years of follow-up. Of these couples, 569 became pregnant without treatment and 863 received IUI treatment. Cumulative probabilities over time of pregnancy without treatment and of IUI treatment can be found in Appendix B. Of the 863 couples who

received treatment, 163 became pregnant after treatment.

We set time 0 at the completion of workup, marking the start of "expectant management" (starting state, state 1). Some couples successfully achieved pregnancy (recovery, state 3) within 1.5 years, whereas others remained in either state 1 (they did not start treatment and did not conceive within 1.5 years) or 2 (they started treatment but did not conceive within 1.5 years). Treatment initiation could occur at any time during follow-up, making it a continuous variable.

The estimands of interest were the probability of getting pregnant by 1.5 years if treatment was: (i) initiated at 0 months; (iii) initiated at 6 months; (iii) not initiated within 1.5 years. We included in the model the baseline covariates which were tested during the couples' workup and which may influence the future pregnancy probability as well as the choice to start IUI quickly or delay its initiation. These covariates are female age, subfertility duration, gynecologist referral (yes/no), infertility type (primary = no pregnancies before/ secondary = lasting pregnancy before), fallopian tubal blockage (no blockage, 1-sided blockage, no test), percentage of progressive sperm count. The covariates age and progressive sperm count were centered and standardized for the analysis.

We modeled the two outcome transitions $1 \rightarrow 3$ and $2 \rightarrow 3$ by means of a Cox proportional hazards model, assuming proportional hazards and linearity. Confidence intervals were obtained by bootstrapping.

We relied on the assumptions of consistency, positivity, conditional exchangeability, conditional independent censoring and correct model specification. While the assumptions of consistency and conditional exchangeability are not testable, they were plausible in our context. The well-defined delay periods supported the assumption of consistency, and the fact that no further evaluations were performed on these couples after workup completion supported the assumption of conditional exchangeability. We evaluated the validity of the positivity assumption in the data. Details can be found in Appendix B. The assumption of a conditionally independent censoring mechanism, although it could not be checked in the data, also appeared plausible. In our analysis, couples were mainly censored to start in vitro fertilization (IVF). The decision to start with IVF was largely influenced by the covariates in the dataset, and thus conditional independence seemed plausible to assume [35]. For the correct model specification assumption in the multi state model, we checked the assumptions of linearity (by plotting martingale residuals) and proportional hazards (by testing the Schoenfeld residuals) for both transitions $1 \rightarrow 3$ and $2 \rightarrow 3$. Based on the linearity checks, subfertility duration and treatment delay were log-transformed before being used as covariates. No notable violations of the proportional hazards assumption were found. We assessed the completeness of follow-up for the study population over 1.5 years for both transitions $1 \rightarrow 3$ and $2 \rightarrow 3$ by means of a reverse Kaplan-Meier, to ensure that we had an adequate number of individuals in the at-risk set to at all time points. Figures and further details can be found in Appendix B.

Figure 6.3 shows the results of our analysis. Our findings suggest that, on average, initiating treatment directly or delaying treatment for 6 months yields similar pregnancy probabilities at 1.5 years post-workup. This implies that delaying treatment does not significantly diminish the likelihood of pregnancy at 1.5 years, while reducing the number of couples undergoing treatment. Not initiating IUI treatment before 1.5 years leads to lower pregnancy probabilities

at 1.5 years. However, the wide confidence intervals prevent us from firm conclusions. For the "treat at 0 months" strategy, confidence intervals are particularly wide: this aligns with transition $2 \rightarrow 3$ having only few subjects that were followed for 1.5 years.

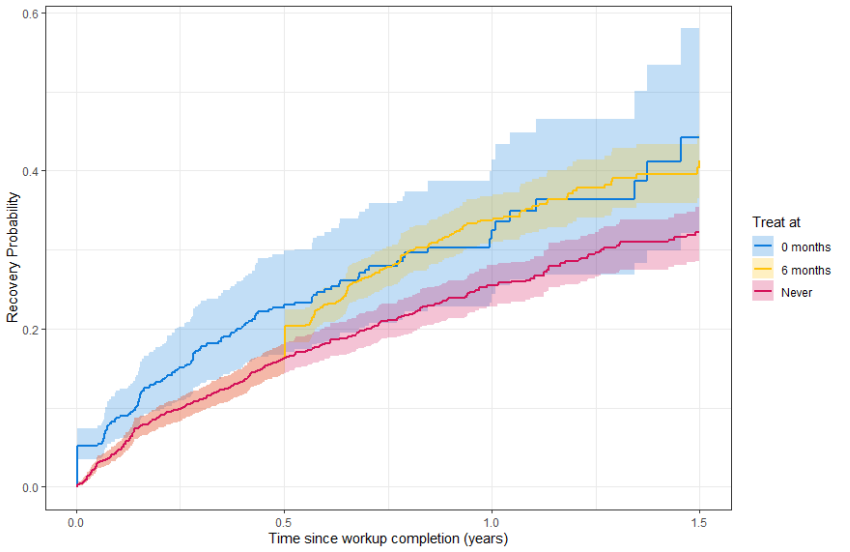

Figure 6.3: Estimated cumulative pregnancy probability for unexplained subfertile couples if they are assigned to different treatment strategies.

6.5 Discussion

In this paper, we propose an illness-death model combined with g-computation that allows estimating the impact of treatment delay in observational data where all patients commence without treatment, and the intended treatment delay remains unobserved for patients who recover before starting treatment. Illness-death modeling provides a natural framework for this semi-competing risks problem. The strength of our work lies in the careful formulation of the set of assumptions under which it is possible to use an illness-death model to draw causal conclusions on the effect of treatment delay. As a key finding, we demonstrated that the identifiability conditions commonly used in causal inference - consistency, positivity, and conditional exchangeability - imply, in the presented illness-death model, that other transition rates remain unchanged after modifying the transition to treatment. While survival analysis experts, as noted in the introduction, often caution against assuming that other transition rates are unaffected by such modifications, we formally show that, under these identifiability conditions, the transition rates for $1 \rightarrow 3$ and $2 \rightarrow 3$ indeed remain unchanged when the transition $1 \rightarrow 2$ is modified, as detailed in Section 6.2.4. The identifiability conditions provide a more intuitive framework compared to directly assuming that transition rates remain unchanged.

In our proposed estimation approach, we reset the clock after treatment and include treatment delay as a covariate. When using illness-death models, researchers can generally choose between a *clock-forward* or *clock-reset* approach, corresponding to Markov and Markov renewal models, respectively, and decide whether to include the time of arrival in the state as a covariate, which further relaxes the Markov (renewal) assumption. As remarked in the paper by Putter et al. [2], these modeling choices should primarily be informed by the clinical context. Our modeling choice was motivated by two main considerations. First, in our data application the more relevant time scale after treatment is time since treatment. Second, data may be insufficient for a correct estimation of the hazard of transition $2 \rightarrow 3$ with a clock-forward approach, as noted in Section 6.2.5. It is important to highlight that transition $2 \rightarrow 3$ is the only transition for which these modeling choices require careful evaluation. The other two transitions, which create a competing risk scenario, are always Markovian due to the absence of prior event history [2].

With simulations we compared the performance of our proposed method to the clone-censor-reweight method. Both approaches are expected to provide asymptotically unbiased estimates of the probability of recovery under different treatment strategies when their respective assumptions are met. We evaluated scenarios in which the modeling assumptions for both methods held, as well as cases where these assumptions were violated, illustrating where and how each method may fail in limited sample sizes. Our proposed method may fail when either one (or both) of the transitions to the outcome is misspecified, whereas the clone-censor-reweight method may fail when the time-to-treatment model for the weights is misspecified. In the scenarios where the modeling assumptions for both methods hold, our multistate model exhibits greater efficiency (smaller variance) in utilizing data compared to the clone-censor-reweight approach, as it borrows information across different delay strategies for both the effect of the covariates and the effect of treatment delay. This finding is consistent with existing literature, where g-computation typically outperforms inverse probability weighting methods in terms of efficiency [36–38]. Double robust methods are able to provide consistent estimates when either the model for the weights or the outcome model is correctly specified. Exploring double-robust methods for estimating treatment delay could be a promising avenue for future research.

A limitation of our method is that it does not account for time-varying confounding, which occurs in situations where treatment decisions are influenced by prognostic factors beyond baseline characteristics. This happens, for instance, if patients are monitored regularly and treatment decisions are made based on their current heath status. Our data application provided an ideal setting to demonstrate our methodology, as treatment initiation decisions were made using only baseline information. Exploring extensions of the current methodology to incorporate time-varying confounders could be a potentially valuable direction for future research.

A second limitation is that our method does not account for competing events, which are relevant when patients can reach a state where treatment is no longer an option. For example, if treatment time was initially planned but the patient later developed conditions preventing them from receiving it, the current approach would need adaptation. In our data application no such events occurred. While death and treatment ineligibility could in principle be competing events, their probability within our target population (couples trying to conceive with unex-

plained subfertility) was negligible, with no occurrences in our dataset. A possible extension of our approach could involve defining a new treatment strategy of interest, such as "treat at t_g if not yet recovered and no competing event has occurred", with the cumulative incidence of recovery if everyone followed treatment strategy g as the estimand. Extending identifiability conditions to this scenario warrants further exploration.

Acknowledgments

We would like to thank Rik van Eekelen, Kristine Openshaw and Matea Skypala for their contribution in this project. We also thank the CECERM study group (Collaborative Effort for Clinical Evaluation in Reproductive Medicine) who collected the data with grant support from ZonMw, The Netherlands Organization for Health Research and Development, The Hague, The Netherlands, grant 945/12/002. Finally, we would like to thank the anonymous reviewers and the associate editor for their comments and suggestions.

Supplementary Material

The Supplementary Material for this article can be found online at: `https://doi.org/10.1002/sim.70061`.
The data and R code used for this paper can be found on Github at `https://github.com/survival-lumc/CausalMultistate`.

References

1. Meira-Machado L, de Uña-Álvarez J, Cadarso-Suárez C, Andersen PK. Multi-State Models for the Analysis of Time-to-Event Data. *Stat Methods Med Res.* 2009;18(2):195–222.
2. Putter H, Fiocco M, Geskus RB. Tutorial in Biostatistics: Competing Risks and Multi-State Models. *Stat Med.* 2007;26(11):2389–2430.
3. Andersen PK, Pohar Perme M. Inference for Outcome Probabilities in Multi-State Models. *Lifetime Data Analysis.* 2008;14(4):405–431.
4. Tsiatis A. A Nonidentifiability Aspect of the Problem of Competing Risks. *Proceedings of the National Academy of Sciences.* 1975;72(1):20–22.
5. Prentice RL, Kalbfleisch JD, Peterson AV, Flournoy N, Farewell VT, Breslow NE. The Analysis of Failure Times in the Presence of Competing Risks. *Biometrics.* 1978;34(4):541.
6. Cox DR. The Analysis of Exponentially Distributed Life-Times with Two Types of Failure. *J R Stat Soc Series B Stat Methodol.* 1959;21(2):411–421.
7. Moeschberger ML, David HA. Life Tests under Competing Causes of Failure and the Theory of Competing Risks. *Biometrics.* 1971;27(4):909–933.
8. Gail M. A Review and Critique of Some Models Used in Competing Risk Analysis. *Biometrics.* 1975;31(1):209–222.
9. Keiding N, Klein JP, Horowitz MM. Multi-State Models and Outcome Prediction in Bone Marrow Transplantation. *Stat Med.* 2001;20(12):1871–1885.
10. Pepe MS, Mori M. Kaplan—Meier, Marginal or Conditional Probability Curves in Summarizing Competing Risks Failure Time Data?. *Stat Med.* 1993;12(8):737–751.
11. Andersen PK, Keiding N. Interpretability and Importance of Functionals in Competing Risks and Multistate Models. *Stat Med.* 2012;31(11-12):1074–1088.
12. Hernán MA. The C-Word: Scientific Euphemisms Do Not Improve Causal Inference From Observational Data. *Am J Public Health.* 2018;108(5):616–619.
13. Gran JM, Lie SA, Øyeflaten I, Borgan Ø, Aalen OO. Causal Inference in Multi-State Models–Sickness Absence and Work for 1145 Participants after Work Rehabilitation. *BMC Public Health.* 2015;15(1):1082.
14. Valeri L, Proust-Lima C, Fan W, Chen JT, Jacqmin-Gadda H. A Multistate Approach for the Study of Interventions on an Intermediate Time-to-Event in Health Disparities Research. *Stat Methods Med Res.* 2023;32(8):1445–1460.
15. Young JG, Stensrud MJ, Tchetgen EJT, Hernán MA. A Causal Framework for Classical Statistical Estimands in Failure Time Settings with Competing Events. *Stat Med.* 2020;39(8):1199–1236.
16. Erdmann A, Loos A, Beyersmann J. A Connection between Survival Multistate Models and Causal Inference for External Treatment Interruptions. *Stat Methods Med Res.* 2023;32(2):267–286.
17. Hernán MA. How to Estimate the Effect of Treatment Duration on Survival Outcomes Using Observational Data. *BMJ.* 2018;360:k182.
18. Maringe C, Benitez Majano S, Exarchakou A, *et al.* Reflection on Modern Methods: Trial Emulation in the Presence of Immortal-Time Bias. Assessing the Benefit of Major Surgery for Elderly Lung Cancer Patients Using Observational Data. *Int J Epidemiol.* 2020;49(5):1719–1729.
19. Hernán M, Robins JM. *Causal Inference*. Boca Raton: Chapman & Hall/CRC; 2023.
20. Klein JP, Moeschberger ML. *Survival Analysis: Techniques for Censored and Truncated Data*. New York Berlin Heidelberg: Springer; 2005.
21. Morris TP, White IR, Crowther MJ. Using Simulation Studies to Evaluate Statistical Methods. *Stat Med.* 2019;38(11):2074–2102.
22. Gaber CE, Hanson KA, Kim S, Lund JL, Lee TA, Murray EJ. The Clone-Censor-Weight Method in Pharmacoepidemiologic Research: Foundations and Methodological Implementation. *Curr Epidemiol Rep.* 2024.
23. RStudio Team. *RStudio: Integrated Development Environment for R*. RStudio, PBC, Boston, MA 2020.
24. de Wreede LC, Fiocco M, Putter H. The Mstate Package for Estimation and Prediction in Non- and Semi-Parametric Multi-State and Competing Risks Models. *Comput Methods Programs Biomed.* 2010;99(3):261–274.
25. Therneau TM. *A Package for Survival Analysis in R* 2023. R package version 3.5-7.
26. Wickham H, Averick M, Bryan J, *et al.* Welcome to the Tidyverse. *J Open Source Softw.* 2019;4(43):1686.
27. Bengtsson H. *matrixStats: Functions that Apply to Rows and Columns of Matrices (and to Vectors)* 2023. R package version 1.0.0.

28. Habbema JDF, Collins J, Leridon H, Evers JLH, Lunenfeld B, te Velde ER. Towards Less Confusing Terminology in Reproductive Medicine: A Proposal. *Hum Reprod.* 2004;19(7):1497–1501.
29. Gnoth C, Godehardt E, Frank-Herrmann P, Friol K, Tigges J, Freundl G. Definition and Prevalence of Subfertility and Infertility. *Hum Reprod.* 2005;20(5):1144–1147.
30. Farquhar CM, Liu E, Armstrong S, Arroll N, Lensen S, Brown J. Intrauterine Insemination with Ovarian Stimulation versus Expectant Management for Unexplained Infertility (TUI): A Pragmatic, Open-Label, Randomised, Controlled, Two-Centre Trial. *Lancet.* 2018;391(10119):441–450.
31. Farhi J, Orvieto R. Cumulative Clinical Pregnancy Rates after COH and IUI in Subfertile Couples. *Gynecol Endocrinol.* 2010;26(7):500–504.
32. van Eekelen R, van Geloven N, van Wely M, *et al.* Is IUI with Ovarian Stimulation Effective in Couples with Unexplained Subfertility?. *Hum Reprod.* 2019;34(1):84–91.
33. van der Steeg J, Steures P, Eijkemans M, *et al.* Pregnancy Is Predictable: A Large-Scale Prospective External Validation of the Prediction of Spontaneous Pregnancy in Subfertile Couples. *Hum Reprod.* 2007;22(2):536–542.
34. Custers IM, Steures P, van der Steeg JW, *et al.* External Validation of a Prediction Model for an Ongoing Pregnancy after Intrauterine Insemination. *Fertil Steril.* 2007;88(2):425–431.
35. Van Geloven N, Geskus RB, Mol BW, Zwinderman AH. Correcting for the Dependent Competing Risk of Treatment Using Inverse Probability of Censoring Weighting and Copulas in the Estimation of Natural Conception Chances. *Stat Med.* 2014;33(26):4671–4680.
36. Goetghebeur E, le Cessie S, De Stavola B, Moodie EE, Waernbaum I. Formulating Causal Questions and Principled Statistical Answers. *Stat Med.* 2020;39(30):4922–4948.
37. Denz R, Klaaßen-Mielke R, Timmesfeld N. A Comparison of Different Methods to Adjust Survival Curves for Confounders. *Stat Med.* 2023;42(10):1461–1479.
38. Ren J, Cislo P, Cappelleri JC, Hlavacek P, DiBonaventura M. Comparing G-Computation, Propensity Score-Based Weighting, and Targeted Maximum Likelihood Estimation for Analyzing Externally Controlled Trials with Both Measured and Unmeasured Confounders: A Simulation Study. *BMC Med Res Methodol.* 2023;23(1):18.