A Novel Approach to the Estimation of Causal Effects of Multiple Event Point Interventions in Continuous Time

Johan Sebastian Ohlendorff[1], **Anders Munch**[1], and **Thomas Alexander Gerds**[1]

[1] University of Copenhagen, Section of Biostatistics

Abstract

In medical research, causal effects of treatments that may change over time on an outcome can be defined in the context of an emulated target trial. We are concerned with estimands that are defined as contrasts of the absolute risk that an outcome event occurs before a given time horizon τ under prespecified treatment regimens. Most of the existing estimators based on observational data require a projection onto a discretized time scale Rose & van der Laan (2011). We consider a recently developed continuous-time approach to causal inference in this setting Rytgaard et al. (2022), which theoretically allows preservation of the precise event timing on a subject level. Working on a continuous-time scale may improve the predictive accuracy and reduce the loss of information. However, continuous-time extensions of the standard estimators comes at the cost of increased computational burden. We will discuss a new sequential regression type estimator for the continuous-time framework which estimates the nuisance parameter models by backtracking through the number of events. This estimator significantly reduces the computational complexity and allows for efficient, single-step targeting using machine learning methods from survival analysis and point processes, enabling robust continuous-time causal effect estimation.

1. Introduction

In medical research, the estimation of causal effects of treatments over time is often of interest. We consider a longitudinal continuous-time setting that is very similar to Rytgaard et al. (2022) in which patient characteristics can change at subject-specific times. This is the typical setting of registry data, which usually contains precise information about when events occur, e.g., information about drug purchase history, hospital visits, and laboratory measurements. This approach offers an advantage over discretized methods, as it eliminates the need to select a time grid mesh for discretization, which can affect both the bias and variance of the resulting estimator. A continuous-time approach would adapt to the events in the data. Furthermore, continuous-time data captures more precise information about when events occur, which may be valuable in a predictive sense. Let τ_{end} be the end of the observation period. We will focus on the estimation of the interventional cumulative incidence function in the presence of time-varying confounding at a specified time horizon $\tau < \tau_{\text{end}}$.

Assumption 1 (Bounded number of events): In the time interval $[0, \tau_{\text{end}}]$ there are at most $K - 1 < \infty$ many changes of treatment and covariates in total for a single individual. Without loss of register data applications, we assume that the maximum number of treatment and covariate changes of an individual is bounded by $K = 10,000$. Practically, we shall adapt K to our data and our target parameter. We let $K - 1$ be given by the maximum number of non-terminal events for any individual in the data.

Let $\kappa_i(\tau)$ be the number of events for individual i up to time τ. In Rytgaard et al. (2022), the authors propose a continuous-time LTMLE for the estimation of causal effects in which a single step of the targeting procedure must update each of the nuisance estimators $\sum_{i=1}^{n} \kappa_i(\tau)$ times. We propose an estimator where the number of nuisance parameters is reduced to $\sim \max_i \kappa_i(\tau)$ in total, and, in principle, only one step of the targeting procedure is needed to update all nuisance parameters. We provide an iterative conditional expectation formula that, like Rytgaard et al. (2022), iteratively updates the nuisance parameters. The key difference is that the estimation of the nuisance parameters can be performed by going back in the number of events instead of going back in time. The different approaches are illustrated in Figure 2 and Figure 3 for an outcome Y of interest. Moreover, we argue that the nuisance components can be estimated with existing machine learning algorithms from the survival analysis and point process literature. As always let $(\Omega, \mathcal{F}, P)$ be a probability space on which all processes and random variables are defined.

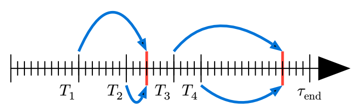

Figure 1: The "usual" approach where time is discretized. Each event time and its corresponding mark is rolled forward to the next time grid point, that is the values of the observations are updated based on the on the events occuring in the previous time interval.

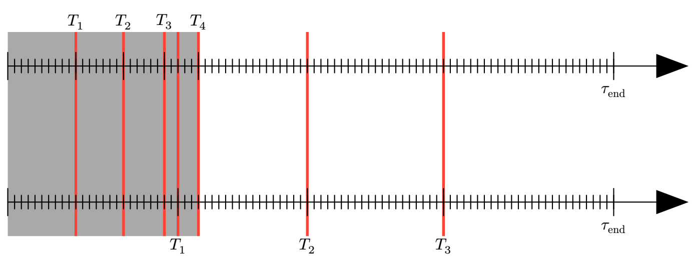

Figure 2: The figure illustrates the sequential regression approach given in Rytgaard et al. (2022) for two observations: Let $t_1 < \ldots < t_m$ be all the event times in the sample. Then, given $\mathbb{E}_Q\left[Y \mid \mathcal{F}_{t_r}\right]$, we regress back to $\mathbb{E}_Q\left[Y \mid \mathcal{F}_{t_{r-1}}\right]$ (through multiple regressions).

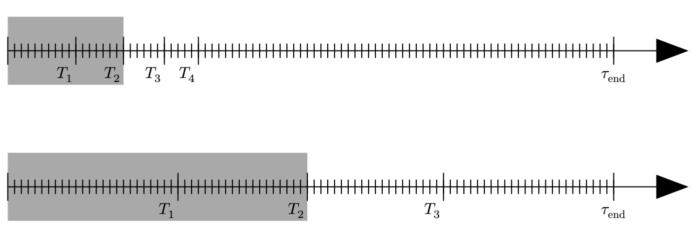

Figure 3: The figure illustrates the sequential regression approach proposed in this article. For each event k in the sample, we regress back on the history $\mathcal{F}_{T_{(k-1)}}$. That is, given $\mathbb{E}_Q\left[Y \mid \mathcal{F}_{T_{(k)}}\right]$, we regress back to $\mathbb{E}_Q\left[Y \mid \mathcal{F}_{T_{(k-1)}}\right]$. In the figure, $k = 3$.

First, we assume that at baseline, we observe the treatment A_0 and the time-varying covariates at time 0, L_0. We let $\pi_0(\cdot \mid L_0)$ be the density of the baseline treatment given the covariates at time 0 (with respect to some measure ν_a) and μ_0 be the density of time-varying covariates at time 0 (with respect to some measure ν_ℓ).

Let N^a be a counting process generating random times at which treatment may change and denote these times by $T_1^a < T_2^a < \ldots < T^a_{N^a(\tau_{\text{end}})}$. Let $A(t) \in \mathcal{A}$ be a càdlàg treatment process defined with values in a finite set of treatment options such that $\Delta A(t) = A(t) - A(t-) = 0$ for all $t \in [0, \tau_{\text{end}}]$ with $\Delta N^a(t) = 0$. Similarly, let N^ℓ be a counting process generating random times at which the covariates may change and denote these times by $T_1^\ell < T_2^\ell < \ldots < T^\ell_{N^\ell(\tau_{\text{end}})}$. Let $L(t) \in \mathcal{L}$ be a càdlàg covariate process with values in a finite subset of $\mathbb{R}^d$ defined on the same time interval that only changes at the times $T_1^\ell < T_2^\ell < \ldots < T^\ell_{N^\ell(\tau_{\text{end}})}$. We assume that the jump times differ with probability 1. Moreover, we assume that only a bounded number of events occur for each individual in the time interval $[0, \tau_{\text{end}}]$ (Assumption 1). We are interested in the cumulative incidence function, so we also observe N^y and N^d corresponding to the counting processes for the primary and competing event, respectively. Finally, let N^c be the counting process for the censoring counting process. Our the outcome of interest is $Y_\tau = I(T \leq \tau, \Delta = y)$, where T is the time of the terminal event and $\Delta \in \{c, y, d\}$ is the indicator for which terminal event occurred. Furthermore, let $T_{(k)}$ be the time of the k'th ordered event in the sample, and define the counting process $N_t = \sum_{k=1}^K I\left(T_{(k)} \leq t\right)$ for the events in the sample. Then we define, for for Borel measurable $\mathbb{A}$ and $\mathbb{L}$, the counting process measures,

$$
\begin{aligned}
N_t^\ell(\mathbb{L}) &= \sum_{i=1}^{N^\ell(\tau_{\text{end}})} I\left(T_i^\ell \leq t, L\left(T_i^\ell\right) \in \mathbb{L}\right) \\
N_t^a(\mathbb{A}) &= \sum_{i=1}^{N^a(\tau_{\text{end}})} I\left(T_i^a \leq t, A(T_i^a) \in \mathbb{A}\right)
\end{aligned}
$$

Though the assumptions so far are stated for stochastic processes, we will instead work with random variables that take values in Euclidean spaces. This is because the processes are assumed to be càdlàg, so that the information about them in the interval $[0, \tau_{\text{end}}]$ can be represented by a finite number of random variables (Assumption 1). Specifically, we claim that each of the observation can be written in the form

$$
O = \left(L_0, A_0, T_{(1)}, D_{(1)}, L\left(T_{(1)}\right), A\left(T_{(1)}\right), \ldots, T_{(K)}, D_{(K)}, L\left(T_{(K)}\right), A\left(T_{(K)}\right)\right).
$$

where $O \sim P$ and the sample is given by n i.i.d. copies of O, i.e., the observations are from a marked point process. This can be seen by defining recursively,

$$
\begin{aligned}
T_{(k)} &= \inf\left\{t > T_{(k-1)} : \Delta N_t^x > 0 \text{ for some } x \in \{a, \ell, c, d, y\}\right\}, \\
D_{(k)} &= \begin{cases} x \text{ if } \Delta N^x_{T_{(k)}} \neq 0 \text{ for } \ x \in \{a, \ell, c, d, y\} \text{ and } T_{(k-1)} < \infty \text{ and } D_{(k-1)} \in \{a, \ell\} \\ \emptyset \text{ otherwise} \end{cases} \\
L\left(T_{(k)}\right) &= \begin{cases} L\left(T_i^\ell\right) \text{ if } D_{(k)} = \ell \text{ and } T_{(k)} = T_i^\ell \text{ for some } \ i \\ \\ L\left(T_{(k-1)}\right) \text{ if } D_{(k)} = a \\ \emptyset \text{ if } D_{(k)} \in \{c, y, d\} \end{cases} \\
A\left(T_{(k)}\right) &= \begin{cases} A(T_i^a) \text{ if } D_{(k)} = a \text{ and } T_{(k)} = T_i^a \text{ for some } \ i \\ \\ A\left(T_{(k-1)}\right) \text{ if } D_{(k)} = \ell \\ \emptyset \text{ if } D_{(k)} \in \{c, y, d\} \end{cases}
\end{aligned}
$$

The history up to the k'th event (and including) can be defined recursively by

$$\mathcal{F}_{T_{(0)}} := \mathcal{F}_0 = \sigma(L_0, A_0)$$

for $k = 0$. For $k > 0$, the history[1] is represented by

$$\mathcal{F}_{T_{(k)}} = \sigma\Big(\Big(T_{(k)}, D_{(k)}, L\Big(T_{(k)}\Big), A\Big(T_{(k)}\Big)\Big)\Big) \vee \mathcal{F}_{T_{(k-1)}},$$

and in fact $\mathcal{F}_{T_{(K)}} = \sigma(O)$.

Theorem 1 shows that if we use the natural filtration, we can take the intensities to be given by Equation 1. This is important as we consider interventions of the intensity of the form given by Equation 1.

[1]Formally, the σ-algebra is not usually defined this way, but because we are using the natural filtration as defined in Theorem 2, the usual definition agrees with this one (see Exercise 4.5.1 of Jacobsen (2006)).

Theorem 1 (Distribution of the events and marks): Let $\pi_{T_{(k)}}\left(\cdot \mid \mathcal{F}_{T_{(k-1)}}\right)$ be the density of the regular conditional distribution of $A\left(T_{(k)}\right)$ given $D_{(k)} = a, T_{(k)}, L\left(T_{(k-1)}\right), A\left(T_{(k-1)}\right), T_{(k-1)} ..., A_0, L_0$ with respect to the measure ν_a. Similarly, let $\mu_t\left(\cdot \mid \mathcal{F}_{T_{(k-1)}}\right)$ be the density of the regular conditional distribution of $L\left(T_{(k)}\right)$ given $D_{(k)} = \ell, T_{(k)}, L\left(T_{(k-1)}\right), A\left(T_{(k-1)}\right), T_{(k-1)} ..., A_0, L_0$ with respect to the measure ν_ℓ. Let $\lambda^x\left(t \mid \mathcal{F}_{T_{(k-1)}}\right)$ be the cause-specific hazard of the x'th $T_{(k)}$ given $L\left(T_{(k-1)}\right), A\left(T_{(k-1)}\right), T_{(k-1)} ..., A_0, L_0$ Then, these are indeed the functions such that

$$
\begin{aligned}
\lambda_t^\ell(\mathbb{L}) &= \lambda^\ell\left(t \mid \mathcal{F}_{T_{(N_{t-})}}\right) \int_{\mathbb{L}} \mu_t\left(x \mid \mathcal{F}_{T_{(N_{t-})}}\right) \nu_\ell(\mathrm{d}x) \\
&= \sum_{k=1}^{K-1} \lambda^\ell\left(t \mid \mathcal{F}_{T_{(k-1)}}\right) \int_{\mathbb{L}} \mu_t\left(x \mid \mathcal{F}_{T_{(k-1)}}\right) \nu_\ell(\mathrm{d}x) I_{\left(T_{(k-1)}, T_{(k)}\right]}(t) \\
\lambda_t^a(\mathbb{A}) &= \lambda^a\left(t \mid \mathcal{F}_{T_{(N_{t-})}}\right) \int_{\mathbb{A}} \pi_t\left(x \mid \mathcal{F}_{T_{(N_{t-})}}\right) \nu_a(\mathrm{d}x) \\
&= \sum_{k=1}^{K-1} \lambda^a\left(t \mid \mathcal{F}_{T_{(k-1)}}\right) \int_{\mathbb{A}} \pi_t\left(x \mid \mathcal{F}_{T_{(k-1)}}\right) \nu_a(\mathrm{d}x) I_{\left(T_{(k-1)}, T_{(k)}\right]}(t) \\
\lambda_t^x &= \lambda^x\left(t \mid \mathcal{F}_{T_{(N_{t-})}}\right) = \sum_{k=1}^{K} \lambda^x\left(t \mid \mathcal{F}_{T_{(k-1)}}\right) I_{\left(T_{(k-1)}, T_{(k)}\right]}(t), x \in \{c, y, d\}
\end{aligned}
\tag{1}
$$

are the $\mathcal{F}_t$-intensities[2] of the respective counting process measures, where $\mathcal{F}_t$ is the filtration generated by all considered processes up to time t, meaning that $\mathcal{F}_t = \sigma\left(\left(N_t^\ell(\mathbb{L}), N_t^a(\mathbb{A}), N_t^y, N_t^d, N_t^c\right) \mid t \geq 0, \mathbb{L} \subseteq \mathcal{L}, \mathbb{A} \subseteq \mathcal{A}\right)$.

Conversely, given the functions in Equation 1, we have for Borel measurable sets $\mathbb{L} \in \mathcal{L} \cup \{\emptyset\}$, and $\mathbb{A} \in \mathcal{A} \cup \{\emptyset\}$, and under appropriate uniform integrability conditions on the intensity measure[3] that

$$
\begin{aligned}
&P\left(T_{(k)} \leq s, D_{(k)} = x, L\left(T_{(k)}\right) \in \mathbb{L}, A\left(T_{(k)}\right) \in \mathbb{A} \mid \mathcal{F}_{T_{(k-1)}}\right) \\
&= \int_0^s \underbrace{\exp\left(- \sum_{x=y,d,\ell,a,c} \int_0^t \lambda^x\left(s \mid \mathcal{F}_{T_{(k-1)}}\right) \mathrm{d}s\right)}_{\text{probability of surviving up to } t.} \underbrace{\lambda^x\left(t \mid \mathcal{F}_{T_{(k)}}\right)}_{\text{probability that it was an event of type } x} \\
&\left(\underbrace{\int_{\mathbb{L}} \mu_t\left(x \mid \mathcal{F}_{T_{(k-1)}}\right) \nu_\ell(\mathrm{d}x)}_{\text{probability of } L\left(T_{(k)}\right) \in \mathbb{L} \text{ given } D_{(k)} = \ell \text{ and } T_{(k)} = t} I\left(x = \ell, A\left(T_{(k-1)}\right) \in \mathbb{A}\right) + \right. \\
&\left. + \underbrace{\int_{\mathbb{A}} \pi_t\left(x \mid \mathcal{F}_{T_{(k-1)}}\right) \nu_a(\mathrm{d}x)}_{\text{probability of } A\left(T_{(k)}\right) \in \mathbb{A} \text{ given } D_{(k)} = a \text{ and } T_{(k)} = t} I\left(x = a, L\left(T_{(k-1)}\right) \in \mathbb{L}\right) + I(x \in \{d, y, c\}, \emptyset \in \mathbb{L}, \emptyset \in \mathbb{A}) \right) \mathrm{d}t.
\end{aligned}
\tag{2}
$$

whenever $D_{(k-1)} \in \{a, \ell\}$ and $T_{(k-1)} < \infty$.

[2] $N_t^x(B) - \int_0^t \lambda_s^x(B) \mathrm{d}s$ is an $\mathcal{F}_t$ martingale for all Borel measurable sets B.

[3] See the Optional Sampling Theorem in Theorem B.0.12 of Jacobsen (2006).

Proof: The theorem is an appropriate extension of Proposition II.7.1 of Andersen et al. (1993) to the multivariate marked point process setting. First note that we can write,

$$N_t^a(\mathbb{A}) = \sum_{k=1}^{K-1} I\left(T_{(k)} \leq t, D_{(k)} = a, A\left(T_{(k)}\right) \in \mathbb{A}\right)$$

which we can extend to

$$N_t^a(\mathbb{L} \times \mathbb{A}) = \sum_{k=1}^{K-1} I\left(T_{(k)} \leq t, D_{(k)} = a, A\left(T_{(k)}\right) \in \mathbb{A}, L\left(T_{(k-1)}\right) \in \mathbb{L}\right)$$

Similarly, we can define $N_t^\ell(\mathbb{L} \times \mathbb{A})$. Moreover, we can write

$$N_t^x(\mathbb{L} \times \mathbb{A}) = \sum_{k=1}^{K} I\left(T_{(k)} \leq t, D_{(k)} = x, \emptyset \in \mathbb{A}, \emptyset \in \mathbb{L}\right)$$

for $x \in \{c, y, d\}$. Define now the counting process

$$N_t(\{x\} \times \mathbb{L} \times \mathbb{A}) = N_t^x(\mathbb{L} \times \mathbb{A})$$

and note that $\mathcal{F}_t = \sigma(N_t(\mathbb{X} \times \mathbb{L} \times \mathbb{A}) \mid t \geq 0, \mathbb{X} \subseteq \{\ell, y, a, d, c\}, \mathbb{L} \subseteq \mathcal{L} \cup \{\emptyset\}, \mathbb{A} \subseteq \mathcal{A} \cup \{\emptyset\})$.

Assume that the conditional distributions of the marks and event times are given as in the theorem. Then an intensity measure is given by Equation 1 by Theorem 4.4.1 of Jacobsen (2006) or section 1.10 of Last & Brandt (1995) or in the counting process setting Proposition II.7.1 of Andersen et al. (1993) for the marked point process N_t.

Conversely, suppose that the intensity measures are given as in Equation 1. Then,

$$\begin{aligned}
&\lambda_t(\mathbb{X} \times \mathbb{L} \times \mathbb{A}) \\
&= \lambda^\ell\left(t \mid \mathcal{F}_{T_{(N_{t-})}}\right) \int_{\mathbb{L}} \mu_t\left(x \mid \mathcal{F}_{T_{(N_{t-})}}\right) \nu_\ell(\mathrm{d}x) I\left(\ell \in \mathbb{X}, A\left(T_{(N_{t-})}\right) \in \mathbb{A}\right) \\
&+ \lambda^a\left(t \mid \mathcal{F}_{T_{(N_{t-})}}\right) \int_{\mathbb{A}} \pi_t\left(x \mid \mathcal{F}_{T_{(N_{t-})}}\right) \nu_a(\mathrm{d}x) I\left(a \in \mathbb{X}, L\left(T_{(N_{t-})}\right) \in \mathbb{L}\right) \\
&+ \lambda^x\left(t \mid \mathcal{F}_{T_{(N_{t-})}}\right) I(d \in \mathbb{X} \vee y \in \mathbb{X} \vee c \in \mathbb{X}, \emptyset \in \mathbb{L}, \emptyset \in \mathbb{A})
\end{aligned} \tag{3}$$

is an intensity measure for the marked point process N_t. This means that the truncated counting measure ${}_{T_{(k-1)}}N_t(\{x\} \times \mathbb{L} \times \mathbb{A}) = N_t(\{x\} \times \mathbb{L} \times \mathbb{A}) - N_{t \wedge T_{(k-1)}}(\{x\} \times \mathbb{L} \times \mathbb{A})$ has the intensity measure

$${}_{T_{(k-1)}}\lambda_t(\{x\} \times \mathbb{L} \times \mathbb{A}) = \lambda_t(\{x\} \times \mathbb{L} \times \mathbb{A}) I_{\left(T_{(k-1)}, \infty\right)}(t).$$

see e.g., Proposition II.4.2 of Andersen et al. (1993) or p. 308 of Jacobsen (2006). Thus,

$$\begin{aligned}
N_t^{(k)}(\{x\} \times \mathbb{L} \times \mathbb{A}) &:= I\left(T_{(k)} \leq t, D_{(k)} = x, L\left(T_{(k)}\right) \in \mathbb{L}, A\left(T_{(k)}\right) \in \mathbb{A}\right) \\
&= {}_{T_{(k-1)}}N_t(\{x\} \times \mathbb{L} \times \mathbb{A}) - {}_{T_{(k)}}N_t(\{x\} \times \mathbb{L} \times \mathbb{A})
\end{aligned}$$

has the intensity measure $\lambda_t(\{x\} \times \mathbb{L} \times \mathbb{A}) I_{\left(T_{(k-1)}, T_{(k)}\right]}(t)$, i.e.,

$$N_t^{(k)}(\{x\} \times \mathbb{L} \times \mathbb{A}) - \int_0^t \lambda_s(\{x\} \times \mathbb{L} \times \mathbb{A}) I_{\left(T_{(k-1)}, T_{(k)}\right]}(t) \mathrm{d}s \tag{4}$$

is an $\mathcal{F}_t$-martingale. By the Optional Sampling Theorem,

$$
\begin{aligned}
\mathbb{E}_P\left[N_t^{(k)}(\{x\}\times\mathbb{L}\times\mathbb{A}) \mid \mathcal{F}_{T_{(k-1)}}\right] &= \mathbb{E}_P\left[N_{t\wedge T_{(k)}}^{(k)}(\{x\}\times\mathbb{L}\times\mathbb{A}) \mid \mathcal{F}_{T_{(k-1)}}\right] \\
&= \mathbb{E}_P\left[\int_0^{t\wedge T_{(k)}} \lambda_s\left(\{x\}\times\mathbb{L}\times\mathbb{A} \mid \mathcal{F}_{T_{(k-1)}}\right) I_{\left(T_{(k-1)},T_{(k)}\right]}(s)\mathrm{d}s \mid \mathcal{F}_{T_{(k-1)}}\right] \\
&= \mathbb{E}_P\left[\int_0^{t} \lambda_s^x\left(\mathbb{L}\times\mathbb{A} \mid \mathcal{F}_{T_{(k-1)}}\right) I_{\left(T_{(k-1)},T_{(k)}\right]}(s)\mathrm{d}s \mid \mathcal{F}_{T_{(k-1)}}\right] \\
&= \int_0^{t} \lambda_s^x\left(\mathbb{L}\times\mathbb{A} \mid \mathcal{F}_{T_{(k-1)}}\right) \mathbb{E}_P\left[I_{\left(T_{(k-1)},T_{(k)}\right]}(s) \mid \mathcal{F}_{T_{(k-1)}}\right]\mathrm{d}s \\
&= \int_0^{t} \lambda_s^x\left(\mathbb{L}\times\mathbb{A} \mid \mathcal{F}_{T_{(k-1)}}\right) P\left(T_{(k)} \geq t \mid \mathcal{F}_{T_{(k-1)}}\right)\mathrm{d}s
\end{aligned}
$$

Here use Equation 3 and that $N_{t-} = k-1$ if $T_{(k-1)} < t \leq T_{(k)}$. Hence,

$$
\begin{aligned}
P\left(T_{(k)} \leq t \mid \mathcal{F}_{T_{(k-1)}}\right) &= \mathbb{E}_P\left[N_t^{(k)}(\{y,c,d,a,l\}\times\mathcal{L}\cup\{\emptyset\}\times\mathcal{A}\cup\{\emptyset\}) \mid \mathcal{F}_{T_{(k-1)}}\right] \\
&= \int_0^t \sum_{x=a,l,y,c,d} \lambda_s^x\left(\mathcal{L}\cup\{\emptyset\}\times\mathcal{A}\cup\{\emptyset\} \mid \mathcal{F}_{T_{(k-1)}}\right) P\left(T_{(k)} \geq t \mid \mathcal{F}_{T_{(k-1)}}\right)\mathrm{d}s \\
&= \int_0^t \sum_{x=a,l,y,c,d} \lambda^x\left(s \mid \mathcal{F}_{T_{(k-1)}}\right) P\left(T_{(k)} \geq t \mid \mathcal{F}_{T_{(k-1)}}\right)\mathrm{d}s
\end{aligned}
$$

yielding $P\left(T_{(k)} > t \mid \mathcal{F}_{T_{(k-1)}}\right) = 1 - \int_0^t \sum_{x=a,l,y,c,d} \lambda^x\left(s \mid \mathcal{F}_{T_{(k-1)}}\right) P\left(T_{(k)} > t \mid \mathcal{F}_{T_{(k-1)}}\right)\mathrm{d}s$ for which $P\left(T_{(k)} > t \mid \mathcal{F}_{T_{(k-1)}}\right) = \exp\left(-\sum_{x=a,l,y,c,d}\int_0^t \lambda^x\left(s \mid \mathcal{F}_{T_{(k-1)}}\right)\mathrm{d}s\right)$ is the only solution. □

1.a. Example for $K = 2$ with $\mathcal{L} = \{0,1\}$ and $\mathcal{A} = \{0,1\}$

In this case, we can represent everything via a multi-state model[4]. We consider the set of states represented by time 0, first treatment visit (a) set to 1, first treatment visit (a) set to 0, first covariate visit (ℓ) set to 1, first covariate visit (ℓ) set to 0, primary event, and competing event. We consider a world without censoring and consider the figure in Figure 4.

If we represent the first state as time 0 and second state as the first treatment visit (a) set to 1. Then the counting process,

$$
N_{12}(t) = I\left(T_{(1)} \leq t, D_{(1)} = a, A\left(T_{(1)}\right) = 1, L\left(T_{(1)}\right) = L_0\right)
$$

has the intensity by Equation 4 given by

$$
\lambda^a(t \mid \mathcal{F}_0)\pi_t(1 \mid \mathcal{F}_0)
$$

with similar expressions for transitions to the other states. Moreover, the probability of making this transition before time t is given by Equation 2.

[4]As far as I know the multi-state model framework does not allow the dependence on more than than the latest (semi-Markov).

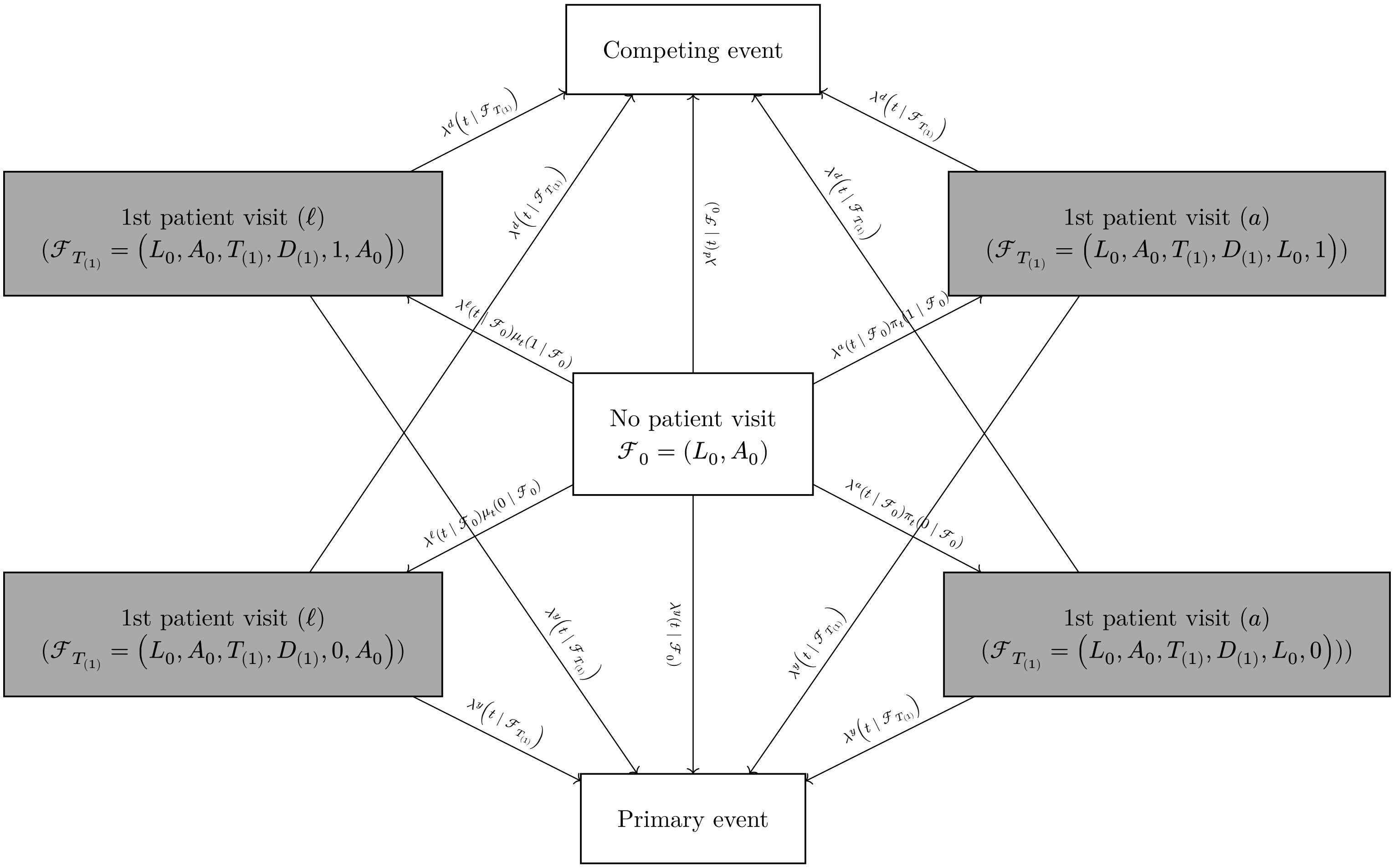

Figure 4: A multi-state model for $K = 2$ with $\mathcal{L} = \mathcal{A} = \{0, 1\}$.

2. A pragmatic approach to continuous-time causal inference

One classical causal inference perspective requires that we know how the data was generated up to unknown parameters (NPSEM) (Pearl, 2009). DAGs are not the most used way of representing the data generating mechanism in the continuous-time setting, but for the event times, we can draw a figure representing the data generating mechanism which is shown in Figure 5.

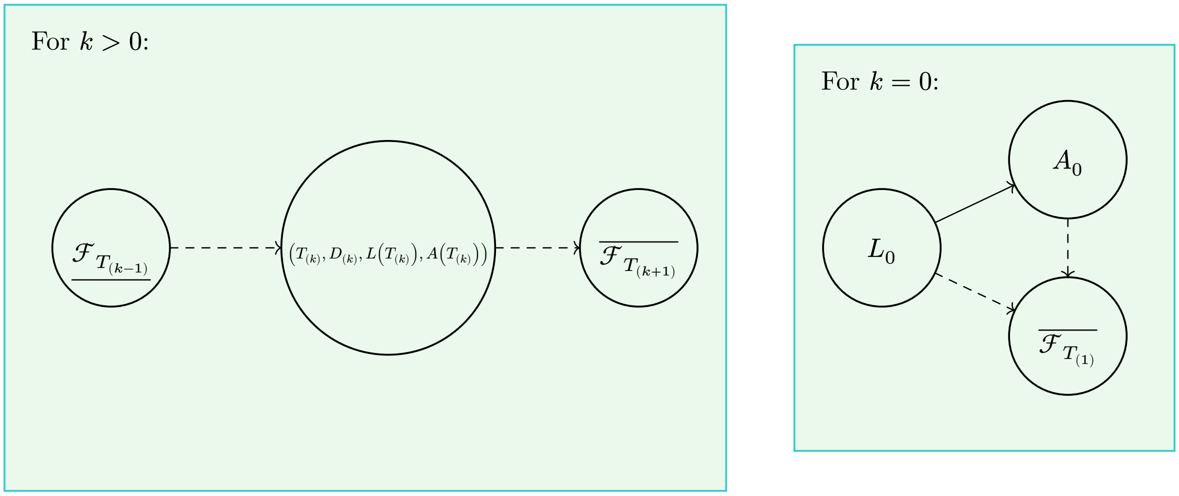

Figure 5: A DAG for the data generating mechanism. The dashed lines indicate multiple edges from the dependencies in the past and into the future. Here $\underline{\mathcal{F}_{T_{(k)}}}$ is the history up to and including the k'th event and $\overline{\mathcal{F}_{T_{(k)}}}$ is the history after and including the k'th event.

Definition 1 (Target parameter): Let Q denote the distribution of an intervention where the random measure λ_t^c and $\pi_t\left(x \mid \mathcal{F}_{T_{(N_{t-})}}\right)$ are replaced by $\lambda_t^{\{c,*\}} = 0$ and $\pi_t^*\left(x \mid \mathcal{F}_{T_{(N_{t-})}}\right)$. We also intervene on the regular conditional distribution of A_0 given L_0 by replacing the density π_0 with π_0^*. Then our target parameter is: $\Psi_\tau(Q) = \mathbb{E}_Q\left[\sum_{k=1}^K I\left(T_{(k)} \le \tau, D_{(k)} = y\right)\right]$.

For this definition, we need a notion of no unobserved confounding, but this is not discussed in the context of continuous-time causal inference (but see e.g., Røysland et al. (2024)). We also need notion of positivity, which we also won't discuss here.

Also note that according to our example with multi-state models with $\mathcal{A} = \mathcal{L} = \{0, 1\}$: If T is the time to the first transition into the primary event or competing event state and D corresponds to the terminal event type, then our target parameter does indeed correspond to the cumulative incidence function at time τ with T and D being the time-to-event and the status, respectively. The target parameter simply summarizes that this can either happen as the first or second event.

We first state and prove a formula for at target parameter that is not causal, but we will use it to identify the causal parameter.

Theorem 2: Let $\bar{Q}_K = I\left(T_{(K)} \le \tau, D_{(K)} = y\right)$ and $\bar{Q}_k = \mathbb{E}_P\left[\sum_{j=k+1}^K I\left(T_{(j)} \le \tau, D_{(j)} = y\right) \mid \mathcal{F}_{T_{(k)}}\right]$. Then,

$$
\begin{aligned}
\bar{Q}_{k-1} = \mathbb{E}_P\Big[&I\left(T_{(k)} \le \tau, D_{(k)} = \ell\right)\bar{Q}_k\left(A\left(T_{(k-1)}\right), L\left(T_{(k)}\right), T_{(k)}, D_{(k)}, \mathcal{F}_{T_{(k-1)}}\right) \\
&+I\left(T_{(k)} \le \tau, D_{(k)} = a\right)\mathbb{E}_P\left[\bar{Q}_k\left(A\left(T_{(k)}\right), L\left(T_{(k-1)}\right), T_{(k)}, D_{(k)}, \mathcal{F}_{T_{(k-1)}}\right) \mid T_{(k)}, D_{(k)}, \mathcal{F}_{T_{(k-1)}}\right] \\
&+I\left(T_{(k)} \le \tau, D_{(k)} = y\right) \Big| \mathcal{F}_{T_{(k-1)}}\Big]
\end{aligned}
$$

for $k = K, ..., 1$. Thus, $\mathbb{E}_P\left[\sum_{k=1}^K I\left(T_{(k)} \le \tau, D_{(k)} = y\right)\right] = \mathbb{E}_P\left[\bar{Q}_0\right]$.

Proof: We find

$$
\begin{aligned}
\bar{Q}_k &= \mathbb{E}_P\left[\sum_{j=k+1}^{K} I\big(T_{(j)} \leq \tau, D_{(j)} = y\big) \mid \mathcal{F}_{T_{(k)}}\right] \\
&= \mathbb{E}_P\left[\mathbb{E}_P\left[\mathbb{E}_P\left[\sum_{j=k+1}^{K} I\big(T_{(j)} \leq \tau, D_{(j)} = y\big) \mid \mathcal{F}_{T_{(k+1)}}\right] \mid T_{(k+1)}, D_{(k+1)}, \mathcal{F}_{T_{(k)}}\right] \mid \mathcal{F}_{T_{(k)}}\right] \\
&= \mathbb{E}_P\Bigg[I\big(T_{(k+1)} \leq \tau, D_{(k+1)} = y\big) \\
&\qquad\qquad + \mathbb{E}_P\left[\mathbb{E}_P\left[\sum_{j=k+2}^{K} I\big(T_{(j)} \leq \tau, D_{(j)} = y\big) \mid \mathcal{F}_{T_{(k+1)}}\right] \mid T_{(k+1)}, D_{(k+1)}, \mathcal{F}_{T_{(k)}}\right] \Bigg| \mathcal{F}_{T_{(k)}}\Bigg] \\
&= \mathbb{E}_P\Bigg[I\big(T_{(k+1)} \leq \tau, D_{(k+1)} = y\big) \\
&+ I\big(T_{(k+1)} \leq \tau, D_{(k+1)} = a\big)\mathbb{E}_P\left[\mathbb{E}_P\left[\sum_{j=k+2}^{K} I\big(T_{(j)} \leq \tau, D_{(j)} = y\big) \mid \mathcal{F}_{T_{(k+1)}}\right] \mid T_{(k+1)}, D_{(k+1)}, \mathcal{F}_{T_{(k)}}\right] \mid \mathcal{F}_{T_{(k)}}\Bigg] \\
&+ \mathbb{E}_P\left[I\big(T_{(k+1)} \leq \tau, D_{(k+1)} = \ell\big)\mathbb{E}_P\left[\sum_{j=k+2}^{K} I\big(T_{(j)} \leq \tau, D_{(j)} = y\big) \mid \mathcal{F}_{T_{(k+1)}}\right] \mid \mathcal{F}_{T_{(k)}}\right]
\end{aligned}
$$

by the law of iterated expectations and that

$$\big(T_{(k)} \leq \tau, D_{(k)} = y\big) \subseteq \big(T_{(j)} \leq \tau, D_{(j)} \in \{a, \ell\}\big)$$

for all $j = 1, ..., k-1$ and $k = 1, ..., K$. $\square$

Theorem 3 (Identification via g-formula): Let $\bar{Q}^a_{k,\tau} = \bar{Q}_k(Q)$ be defined as in the previous theorem for Q. Then, we can identify $\bar{Q}^a_{k,\tau}$ via the intensities as

$$\begin{aligned}
\bar{Q}^a_{k,\tau} = & \int_{T_{(k)}}^{\tau} \exp\left(- \sum_{x \in \{\ell,a,d,y\}} \int_{T_{(k)}}^{s} \lambda^x\left(u \mid \mathcal{F}_{T_{(k)}}\right) \mathrm{d}u\right) \lambda^a\left(s \mid \mathcal{F}_{T_{(k)}}\right) \\
& \qquad \times \left(\int_{\mathcal{A}} \bar{Q}^a_{k+1,\tau}\left(L\left(T_{(k-1)}\right), a_k, s, a, \mathcal{F}_{T_{(k)}}\right) \pi^*_s\left(a_k \mid \mathcal{F}_{T_{(k)}}\right) \nu_A(\mathrm{d}a_k)\right) \mathrm{d}s \\
& + \int_{T_{(k)}}^{\tau} \exp\left(- \sum_{x \in \{\ell,a,d,y\}} \int_{T_{(k)}}^{s} \lambda^x\left(u \mid \mathcal{F}_{T_{(k)}}\right) \mathrm{d}u\right) \lambda^\ell\left(s \mid \mathcal{F}_{T_{(k)}}\right) \\
& \qquad \times \left(\int_{\mathcal{L}} \bar{Q}^a_{k+1,\tau}\left(l_k, A\left(T_{(k-1)}\right), s, \ell, \mathcal{F}_{T_{(k)}}\right) \mu_s\left(l_k \mid \mathcal{F}_{T_{(k)}}\right) \nu_L(\mathrm{d}l_k)\right) \mathrm{d}s \\
& + \int_{T_{(k)}}^{\tau} \exp\left(- \sum_{x \in \{\ell,a,d,y\}} \int_{T_{(k)}}^{s} \lambda^x\left(u \mid \mathcal{F}_{T_{(k)}}\right) \mathrm{d}u\right) \lambda^y\left(s \mid \mathcal{F}_{T_{(k)}}\right) \mathrm{d}s
\end{aligned} \tag{5}$$

Alternatively, we can apply inverse probability of censoring weighting to obtain

$$\begin{aligned}
\bar{Q}^a_{k-1,\tau} = \mathbb{E}_P \Bigg[& \frac{I\left(T_{(k)} \le \tau, D_{(k)} = \ell\right)}{\exp\left(- \int_{T_{(k-1)}}^{T_{(k)}} \lambda^c\left(s \mid \mathcal{F}_{T_{(k)}}\right) \mathrm{d}s\right)} \bar{Q}^a_{k,\tau}\left(A\left(T_{(k-1)}\right), L\left(T_{(k)}\right), T_{(k)}, D_{(k)}, \mathcal{F}_{T_{(k-1)}}\right) \\
& + \frac{I\left(T_{(k)} \le \tau, D_{(k)} = a\right)}{\exp\left(- \int_{T_{(k-1)}}^{T_{(k)}} \lambda^c\left(s \mid \mathcal{F}_{T_{(k)}}\right) \mathrm{d}s\right)} \\
& \qquad \times \int \bar{Q}^a_{k,\tau}\left(a_k, L\left(T_{(k-1)}\right), T_{(k)}, D_{(k)}, \mathcal{F}_{T_{(k-1)}}\right) \pi^*_{T_{(k)}}\left(a_k \mid \mathcal{F}_{T_{(k)}}\right) \nu_A(\mathrm{d}a_k) \\
& + \frac{I\left(T_{(k)} \le \tau, D_{(k)} = y\right)}{\exp\left(- \int_{T_{(k-1)}}^{T_{(k)}} \lambda^c\left(s \mid \mathcal{F}_{T_{(k)}}\right) \mathrm{d}s\right)} \,\Bigg|\, \mathcal{F}_{T_{(k-1)}} \Bigg]
\end{aligned} \tag{6}$$

for $k = K-1, ..., 1$. Then,

$$\Psi_\tau(Q) = \mathbb{E}_P\left[\int \bar{Q}^a_{0,\tau}(a, L_0) \nu_A(\mathrm{d}a)\right].$$

Proof: The theorem is an immediate consequence of Theorem 1 and Theorem 2 (the sets $\left(T_{(k)} \le t, D_{(k)} = x, L\left(T_{(k)}\right) \in \mathbb{L}, A\left(T_{(k)}\right) \in \mathbb{A}\right)$ fully determine the regular conditional distribution of $\left(T_{(k)}, D_{(k)}, L\left(T_{(k)}\right), A\left(T_{(k)}\right)\right)$ given $\mathcal{F}_{T_{(k-1)}}$). □

Interestingly, Equation 5 corresponds exactly with the target parameter of Rytgaard et al. (2022) and Gill & Robins (2023) by plugging in the definitions of $\bar{Q}^a_{k,\tau}$ and simplifying (to be shown).

3. Implementation of the Iterative Conditional Expectations formula

We assume that K_τ is the 1 + the maximal number of non-terminal events that occur before time τ. For now, we assume that this is number is fixed and does not depend on the sample. Let $\tilde{Y}_k(t) = I\left(T_{(k-1)} < t \leq T_{(k)}\right)$.

For $k = K_\tau - 1, ..., 0$:

- We want a prediction function $\bar{Q}^a_{k,\tau}$ of the history up to the k'th event, that is $\bar{Q}^a_{k,\tau} : \mathcal{H}_k \longrightarrow \mathbb{R}$, given that we have one for the $(k+1)$'th event, i.e., $\bar{Q}^a_{k+1,\tau} : \mathcal{H}_k \longrightarrow \mathbb{R}$ (note that for $k = K_\tau$, we have $\bar{Q}^a_{K_\tau,\tau} = I\left(T_{(K_\tau)} \leq \tau, D_{(K_\tau)} = y\right)$.). We consider the data set $\mathcal{D}_{k,n}$ that is obtained from the original data $\mathcal{D}_n$ by only considering the observations that have had k non-terminal events, that is $D_{(k)} \in \{a, \ell\}$ for $j = 1, ..., k$. On this data:
  - We estimate $\lambda^c\left(\cdot \mid \mathcal{F}_{T_{(k+1)}}\right)$ by using $T_{(k+1)}$ as the time-to-event and $D_{(k+1)}$ as the event indicator on the data set $\mathcal{D}_{k,n}$, regressing on $\mathcal{F}_{T_{(k)}} = \left(L\left(T_{(k)}\right), A\left(T_{(k)}\right), T_{(k)}, D_{(k)}, ..., L_0, A_0\right)$[5]

  We are now able provide estimated values for the integrand in Equation 6. These values are provided on the smaller data set $\mathcal{D}_{k,n,\tau}$ of $\mathcal{D}_{k,n}$ where we only consider the observations with $T_{(k)} \leq \tau$. This is done as follows:

  1. For observations in $\mathcal{D}_{k,n,\tau}$ with $D_{(k+1)} = \ell$ and $T_{(k+1)} \leq \tau$, use the previous function to predict values $\bar{Q}^a_{k+1,\tau}\left(L\left(T_{(k+1)}\right), A\left(T_{(k)}\right), T_{(k+1)}, D_{(k+1)}, \mathcal{F}_{T_{(k)}}\right)$.
  2. For observations in $\mathcal{D}_{k,n,\tau}$ with $D_{(k+1)} = a$ and $T_{(k+1)} \leq \tau$, integrate using the previous function $\int_{\mathcal{A}} \bar{Q}^a_{k+1,\tau}\left(L\left(T_{(k)}\right), a_k, T_{(k+1)}, D_{(k+1)}, \mathcal{F}_{T_{(k)}}\right)\pi^*_{T_{(k+1)}}\left(a_k \mid \mathcal{F}_{T_{(k)}}\right)\nu_A(\mathrm{d}a_k)$. If for example the intervention sets the treatment to 1, then $\int_{\mathcal{A}} \bar{Q}^a_{k+1,\tau}\left(L\left(T_{(k)}\right), a_k, T_{(k+1)}, D_{(k+1)}, \mathcal{F}_{T_{(k)}}\right)\pi^*_{T_{(k)}}\left(1 \mid \mathcal{F}_{T_{(k)}}\right)\nu_A(\mathrm{d}a_k) = \bar{Q}^a_{k+1,\tau}\left(L\left(T_{(k)}\right), 1, T_{(k)}, D_{(k)}, \mathcal{F}_{T_{(k)}}\right)$. This gives predicted values for this group.
  3. For observations in $\mathcal{D}_{k,n,\tau}$ with $D_{(k+1)} = y$ and $T_{(k+1)} \leq \tau$, simply put the values equal to 1.
  4. For all other observations put their values equal to 0.

  For all the observations, divide the corresponding values by estimates of censoring survival function $\exp\left(-\int_{T_{(k)}}^{T_{(k+1)}} \lambda^c\left(s \mid \mathcal{F}_{T_{(k)}}\right)\mathrm{d}s\right)$. We then regress the values on $\mathcal{F}_{T_{(k)}} = \left(L\left(T_{(k)}\right), A\left(T_{(k)}\right), T_{(k)}, D_{(k)}, ..., L_0, A_0\right)$. From this regression, we set $\bar{Q}^a_{k,\tau}$ to be the predicted values of the function from the regression.
- If $k = 0$: We estimate the target parameter via $\mathbb{P}_n\left[\sum_{k=1}^{K_\tau} \bar{Q}^a_{0,\tau}(\cdot, a_0)\nu_A(\mathrm{d}a_0)\right]$.

Note: The $\bar{Q}^a_{k,\tau}$ have the interpretation of the heterogenous causal effect after k events.

For now, we recommend Equation 6 for estimating $\bar{Q}^a_{k,\tau}a$: For estimators of the hazard that are piecewise constant, we would need to compute integrals for each unique pair of history and event times occurring in the sample at each event k. On the other hand, the IPCW approach is very sensitive to the specification of the censoring distribution. Something very similar can be written down when we use Equation 5.

3.a. Alternative nuisance parameter estimators

An alternative is to estimate the entire cumulative hazards Λ^x at once instead of having K separate parameters: There are very few methods for marked point process estimation but see Liguori et al. (2023) for methods mostly based on neural networks or Weiss & Page (2013) for a forest-based method. As a final alternative, we can use temporal difference learning to avoid iterative estimation of $\bar{Q}^a, \tilde{Q}$ Shirakawa et al. (2024). Most point process estimators are actually on the form given in terms of Equation 1.

4. The efficient influence function (needs to be updated)

[5] We abuse the notation a bit by writing $\mathcal{F}_{T_{(k)}}$ here, but it is actually a σ-algebra.

We want to use machine learning estimators of the nuisance parameters, so to get inference we need to debias our estimate with the efficient influence function, e.g., double/debiased machine learning Chernozhukov et al. (2018) or targeted minimum loss estimation van der Laan & Rubin (2006). We use Equation 6 for censoring to derive the efficient influence function, because it will contain fewer martingale terms. Let $N^c_k(t) = N_t(\{c\} \times \mathcal{L} \cup \{\emptyset\} \times \mathcal{A} \cup \{\emptyset\})$.

Theorem 4 (Efficient influence function): Let $\bar{Q}^a_{k,\tau}(u) := \frac{I\left(\tilde{T}_{(k)} \leq u, \tilde{D}_{(k)} \neq c\right)}{\exp\left(-\int_{T_{(k-1)}}^{T_{(k)}} \lambda^c\left(s \mid \mathcal{F}_{T_{(k)}}\right) \mathrm{d}s\right)} \tilde{Q}^a_{k+1}\left(T_{(k+1)}, D_{(k+1)}, \mathcal{F}_{T_{(k)}}\right)$ for $u \leq \tau$. The efficient influence function is given by

$$
\begin{aligned}
\varphi^*_\tau(P) = \sum_{k=1}^{K} \prod_{j=0}^{k-1} & \left(\frac{\pi^*_{T_{(j)}}\left(A\left(T_{(j)}\right) \mid \mathcal{F}_{T_{(j-1)}}\right)}{\pi_{T_{(j)}}\left(A\left(T_{(j)}\right) \mid \mathcal{F}_{T_{(j-1)}}\right)} \right)^{I\left(D_{(j)} = a\right)} \frac{I\left(D_{(k-1)} \in \{\ell, a\}, T_{(k-1)} \leq \tau\right)}{\exp\left(-\sum_{1 \leq j < k} \int_{T_{(j-1)}}^{T_{(j)}} \lambda^c\left(s \mid \mathcal{F}_{T_{(j-1)}}\right) \mathrm{d}s\right)} \\
& \times \Bigg[I(k < K) \frac{I\left(D_{(k)} \in \{\ell, a\}, T_{(k)} \leq \tau\right)}{\exp\left(-\int_{T_{(k-1)}}^{T_{(k)}} \lambda^c\left(s \mid \mathcal{F}_{T_{(j-1)}}\right) \mathrm{d}s\right)} \left[\bar{Q}^a_{k,\tau}\left(\tau, \mathcal{F}_{T_{(k)}}\right) - \tilde{Q}^a_k\left(T_{(k)}, D_{(k)}, \mathcal{F}_{T_{(k-1)}}\right)\right] \\
& \quad + \frac{I\left(D_{(k)} \neq c, T_{(k)} \leq \tau\right)}{\exp\left(-\int_{T_{(k-1)}}^{T_{(k)}} \lambda^c\left(s \mid \mathcal{F}_{T_{(j-1)}}\right) \mathrm{d}s\right)} \tilde{Q}^a_k\left(T_{(k)}, D_{(k)}, \mathcal{F}_{T_{(k-1)}}\right) - \bar{Q}^a_{k-1,\tau}\left(\tau, \mathcal{F}_{T_{(k-1)}}\right) \\
& \quad + \int_{T_{(k-1)}}^{\tau} \left(\bar{Q}^a_{k-1,\tau}\left(\tau, \mathcal{F}_{T_{(k-1)}}\right) - \bar{Q}^a_{k-1,\tau}\left(u, \mathcal{F}_{T_{(k-1)}}\right)\right) \frac{1}{\exp\left(-\int_0^u \sum_{x=a,\ell,d,y,c} \lambda^x\left(s \mid \mathcal{F}_{T_{(k-1)}}\right) \mathrm{d}s\right)} \left(N^c_k(\mathrm{d}s) - \tilde{Y}_{k-1}(s) \lambda^c\left(s \mid \mathcal{F}_{T_{(k-1)}}\right) \mathrm{d}s\right) \Bigg] \\
& + \int \tilde{Q}^a_1(a, L_0) \nu_A(\mathrm{d}a) - \Psi_\tau(P)
\end{aligned}
$$

(we take the empty sum to be zero and define $T_0 = 0$, $D_{(0)} = a$ and $\mathcal{F}_{T_{(-1)}} = L(0)$.)

For now, we recommend using the one step estimator and not the TMLE because the martingales are computationally intensive to estimate. This means that multiple TMLE updates may not be a good idea.

5. Data-adaptive choice of K

In practice, we will want to use K_τ to be equal to 1 + maximum number of non-terminal events up to τ in the sample. It turns out, under the boundedness condition of the number of events, that an estimator that is asymptotically linear with efficient influence function $\varphi^*_\tau(P)(\max_i \kappa_i(\tau))$ is also asymptotically linear with efficient influence function $\varphi^*_\tau(P)(K_\tau)$ where K_τ is the last event point such that $P(\kappa_i(\tau) = K_\tau) > 0$.

Sketch: We want to use $K = K_n = \max_i \kappa_i(\tau)$. If we can do asymptotically and efficient inference for K_n, then we can also do it for a limiting $K_n \leq K$. Assume that the estimator is asymptotically linear with efficient influence function $\varphi^*_\tau(P)(K_n)$. Then by Assumption 1, there exists a K_{lim} which is the last point such that $P(K_n = K_{\text{lim}}) > 0$. Then, K_n converges to K_{lim} (by independence), and moreover, under standard regularity conditions such as strict positivity,

$$
(\mathbb{P}_n - P)(\varphi^*_\tau(P)(K_n) - \varphi^*_\tau(P)(K))
$$

is $o_{P\left(n^{-\frac{1}{2}}\right)}$, so if have asymptotic linearity in terms of $\varphi^*_\tau(P)(K_n)$, then we automatically have it for the original estimator for $\varphi^*_\tau(P)(K_{\text{lim}})$

6. Issues relating to rare patient histories (postponed)

Consider the following table representing example data:

k	0	1	2	3	4	5
$\tilde{Y}_k(\tau)$	10000	8540	5560	2400	200	4
$\Delta A\left(T_{(k)}\right)$	6000	3560	1300	100	2	NA
$\Delta L\left(T_{(k)}\right)$	2540	2000	1100	100	2	NA

6.a. Pooling

Some people have complex histories. There may be very few of these people in the sample, so how do we do estimate the cause-specific hazard for the censoring in, say, the first step? In the artificial data example, there are only 4 people at the last time point.

We propose to pool the regressions across event points: Let us say that we want to estimate the cause-specific hazard for the censoring at event $k+1$ among people who are at risk of being censored at the $k+1$'th event, that is they either had a treatment change or a covariate change at their k event. If this population in the sample is very small, then we could do as follows. We delete the first event for these observations. Then the number of covariates is reduced by one, so we have the same number of covariates as we did for the people who are at risk of having an event at the k'th event. We combine these two data sets into one and regress the cause-specific hazard for the censoring at event "k". This provides a data set with correlated observations, which likely is not biased as we are not interested in variance estimation for parameters appearing in the regression.

To estimate the regression for the time-varying covariates, one could do:

- Not intervene on the last two or three time points, letting certain parts of the data generating mechanism be observational, that is $\pi^*_t\left(\cdot \mid \mathcal{F}_{T_{(j)}}\right) = \pi_t\left(\cdot \mid \mathcal{F}_{T_{(j)}}\right)$ for $j = 4, 5$.
- Another is to make a Markov-like assumption in the interventional world, i.e.,

$$\mathbb{E}_Q\left[\sum_{j=1}^{3} I\left(T_{(j)} \leq \tau, D_{(j)} = y\right) \mid \mathcal{F}_0\right] = \mathbb{E}_Q\left[\sum_{j=6}^{8} I\left(T_{(j)} \leq \tau, D_{(j)} = y\right) \mid \mathcal{F}_{T_{(5)}}\right]$$

  So we separately estimate the target parameter on the left hand side and use it to estimate the one on the right when we need to, pooling the data from the last three events with the data from the first three events.

Doing this adaptively leads to data-adaptive target parameter (Hubbard et al., 2016).

Other possible methods are:

- Use an estimation procedure that is similar to Shirakawa et al. (2024) or use hazards which are estimated all at once.
- Bayesian methods may be useful since they do not have issues with finite sample size. They are also a natural way of dealing with the missing data problem. However, nonparametric Bayesian methods are not (yet) able to deal with a large number of covariates.

6.b. Other ideas

Some other issues are that the covariates are (fairly) high dimensional. This may yield issues with regression-based methods.

- Use Early-stopping cross-validation described as follows: First fit models with no covariates. Then we fit a model with the covariates from the last event. Determine if this is improves model fit via cross-validation and then we move on to the two latest changes and so on. Stop when the model fit does not improve. Theorem 2 of Schuler & van der Laan (2022) states that the convergence rates for an empirical risk minimizer are preserved. CTMLE also does something very similar (van der Laan & Gruber, 2010). This way, we may only select variables that are important in the specification of the treatment and outcome mechanism.

6.c. Topics for further research

Interestingly, $\int \bar{Q}^a_{0,\tau}(a, L_0)\nu_A(\mathrm{d}a)$ is a heterogenous causal effect. Can we estimate heterogenous causal effects in this way?

Time-fixed time-varying treatment could probably be interesting within a register-based study since it may be easier to define treatment in an interval rather than two define on, each time point, if the patient is on the treatment or not.

It may also sometimes be the case that some time-varying covariates are measured regularly instead of at subject-specific times. In this case, we may be able to do something similar to the above.

Bibliography

Andersen, P. K., Borgan, Ø., Gill, R. D., & Keiding, N. (1993). *Statistical Models Based on Counting Processes*. Springer US. https://doi.org/10.1007/978-1-4612-4348-9

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., & Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, *21*(1), C1–C68. https://doi.org/10.1111/ectj.12097

Gill, R. D., & Robins, J. M. (2023,). *Causal Inference for Complex Longitudinal Data: The Continuous Time g-Computation Formula*. https://arxiv.org/abs/math/0409436

Hubbard, A. E., Kherad-Pajouh, S., & van der Laan, M. J. (2016). Statistical inference for data adaptive target parameters. *The International Journal of Biostatistics*, *12*(1), 3–19.

Jacobsen, M. (2006). *Point Process Theory and Applications: Marked Point and Piecewise Deterministic Processes*. Birkhäuser Boston. https://doi.org/10.1007/0-8176-4463-6

Last, G., & Brandt, A. (1995). *Marked Point Processes on the Real Line: The Dynamical Approach*. Springer. https://link.springer.com/book/9780387945477

Liguori, A., Caroprese, L., Minici, M., Veloso, B., Spinnato, F., Nanni, M., Manco, G., & Gama, J. (2023, July). *Modeling Events and Interactions through Temporal Processes – A Survey* (Issue arXiv:2303.06067). arXiv. https://doi.org/10.48550/arXiv.2303.06067

Pearl, J. (2009). *Causality: Models, Reasoning and Inference* (2nd ed.). Cambridge University Press.

Rose, S., & van der Laan, M. J. (2011). Introduction to TMLE. In *Targeted Learning: Causal Inference for Observational and Experimental Data* (pp. 67–82). Springer New York. https://doi.org/10.1007/978-1-4419-9782-1_4

Rytgaard, H. C., Gerds, T. A., & Laan, M. J. van der. (2022). Continuous-Time Targeted Minimum Loss-Based Estimation of Intervention-Specific Mean Outcomes. *The Annals of Statistics*, *50*(5), 2469–2491. https://doi.org/10.1214/21-AOS2114

Røysland, K., C. Ryalen, P., Nygård, M., & Didelez, V. (2024). Graphical criteria for the identification of marginal causal effects in continuous-time survival and event-history analyses. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, qkae56. https://doi.org/10.1093/jrsssb/qkae056

Schuler, A., & van der Laan, M. (2022, May). *The Selectively Adaptive Lasso*.

Shirakawa, T., Li, Y., Wu, Y., Qiu, S., Li, Y., Zhao, M., Iso, H., & van der Laan, M. (2024, April). *Longitudinal Targeted Minimum Loss-based Estimation with Temporal-Difference Heterogeneous Transformer* (Issue arXiv:2404.04399). arXiv.

van der Laan, M. J., & Gruber, S. (2010). Collaborative double robust targeted maximum likelihood estimation. *The International Journal of Biostatistics*, *6*(1).

van der Laan, M. J., & Rubin, D. (2006). Targeted Maximum Likelihood Learning. *The International Journal of Biostatistics*, *2*(1). https://doi.org/10.2202/1557-4679.1043

Weiss, J. C., & Page, D. (2013). Forest-Based Point Process for Event Prediction from Electronic Health Records. In H. Blockeel, K. Kersting, S. Nijssen, & F. Železný (Eds.), *Machine Learning and Knowledge Discovery in Databases: Machine Learning and Knowledge Discovery in Databases*. https://doi.org/10.1007/978-3-642-40994-3_35