Competing Event Models: Autoregressive Generation Under Interventions on When and What to Do

Yoav Wald * Technion
Xiang Gao NYU
Sumit Chopra NYU
Juho Lee KAIST
Rajesh Ranganath NYU

Abstract

Modeling interventions that include timing is important in many domains, such as treatments in healthcare and transactions in finance. However, standard next-token predictors, typically implemented using transformers, do not naturally support altering the timing of specific future events. We develop *Competing Event Models* (CEMs), autoregressive world models for events in irregular time: once trained on observational data, a CEM can simulate trajectories under any intervention that can be carried out in practice, i.e. any policy that decides what to do and when based on the observed history. Drawing from the competing risks literature, CEMs model independent distributions for latent times of each event type, where the earliest realized time determines the next event. This makes interventions on timing straightforward and sampling efficient, and through appropriate design choices, CEMs also model concurrent events. We evaluate CEMs on a cancer tumor volume simulator and on prediction tasks using the MIMIC-IV dataset. Our results show that CEMs accurately recover causal effects of sequential treatments in simulations while surpassing standard transformer baselines in predictive performance on 3 out of 4 real-world medical data tasks.

1 Introduction

Autoregressive models are the de-facto solution for generating sequential data in language (Radford et al., 2018), genetic sequences (Rives et al., 2021), medical records (Steinberg et al., 2021), and more. Modeling interventions with such architectures can be applied to tasks like off-policy evaluation (Uehara et al., 2025) and policy optimization (Yu et al., 2020). Consider modeling medical records and observing the data of a patient suffering from heart failure. A cardiologist may wish to assess the risk of worsening conditions (e.g., experiencing decompensation) or side effects under different treatment policies/regimens. Such causal inference questions are important in medicine (Joshi et al., 2025; van Amsterdam et al., 2022; Chen et al., 2021), and generating potential trajectories under regimens of interest is an attractive way to answer them.

In many scenarios, the timing of treatment is a crucial aspect of treatment policies. For example, we may want to understand how the patient trajectory would change if a treatment was administered a month later or a week earlier than planned. More elaborate questions involve multiple types of events (e.g., treatments and measurements) and dynamic policies, such as: "What if in the future we treat the patient whenever cholesterol levels above a certain threshold are measured?". However, we claim here that common autoregressive models of the observational distribution, e.g., transformers trained for next-event or next-token prediction, do not naturally lend themselves to sampling from interventional distributions that modify the timing of a specific event type. This is because these models directly parameterize the time of the next event, regardless of its type. We address this limitation by augmenting autoregressive models to predict *latent* times for each event type. Our contributions are as follows:

*Correspondence to yoav.wald@technion.ac.il

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

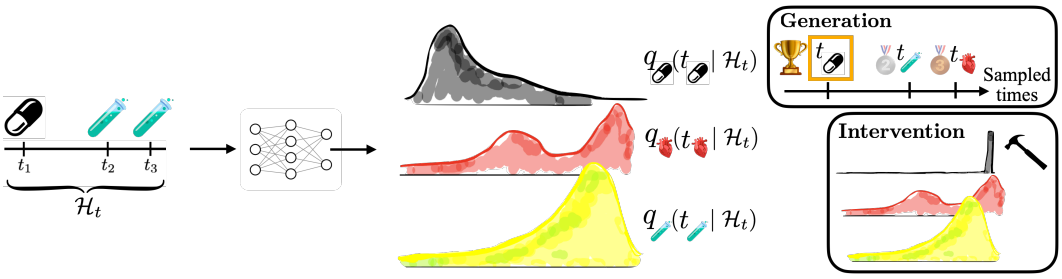


Figure 1: The competing events model for a simplified example with $\mathcal{E} = \{💊, ❤️, 🧪\}$, corresponding to treatment (e.g. statins), lab measurements (cholesterol), and occurrence of cardiovascular disease (myocardial infarction) accordingly. A neural network processes past events $\mathcal{H}_t$ and outputs a time distribution q_e for each type $e \in \mathcal{E}$. The next event type is determined by the smallest sampled time (this is $(t_{🧪}, 🧪)$ in the image). Interventions replace a specific q_e with an alternative distribution.

- We develop Competing Events Models (CEMs), autoregressive models based on a competing-events representation that assigns a latent time variable to each event type. This representation enables interventions on both what events occur and when they occur. The model, sampling procedure, and interventions are illustrated in fig. 1.
- We create a training procedure that scales to large event vocabularies and long trajectories. We also show how to model concurrent events, either directly during training or by converting between the competing-events representation and other representations of the next-event distribution. These capabilities are important for real-world datasets such as MIMIC-IV, which contain many event types, long trajectories, and a large proportion of concurrent events.
- We implement CEMs with transformers and show they enable accurate off-policy evaluation on simulated data and competitive performance on real-world healthcare risk prediction tasks.

2 Modeling Interventions in Point Processes

2.1 The Data Generating Processes

Following notation similar to Upadhyay et al. (2018) and Wald et al. (2025), $\mathcal{E}$ is the set of event types and we add a type END for the end of the sequence, denoting $\overline{\mathcal{E}} = \mathcal{E} \cup \{\text{END}\}$. The time intervals between events take values in a set $\overline{\mathcal{T}} = \mathcal{T} \cup \{\infty\}$, where ∞ is understood as the end of the sequence. The set $\mathcal{T}$ can be continuous $[0, T]$, or a finite set as in the implementation we use in experiments, where time is discretized ($\mathcal{T} = [0, 1, \ldots, T]$, Li et al. (2021)) or binned ($\mathcal{T} = \{[0, b_1], [b_1, b_2], \ldots, [b_k, T]\}$, Renc et al. (2024)). Notation is introduced as needed throughout the paper, and a summary of the notation is in appendix B.

Definition 2.1. P is a distribution over sequences of events in time, where an event is a tuple $(t, e) \in \overline{\mathcal{T}} \times \overline{\mathcal{E}}$. A sequence is $\mathcal{H} = \{(t_k, e_k)\}_{k=1}^{n}$ where only $e_n = \text{END}$, and histories in a sequence are denoted $\mathcal{H}_t := \{(t_k, e_k) \in \mathcal{H} : t_k \leq t\}$. For past events $\mathcal{H}_t$, we denote the density or mass functions of the next event time and type by $P(T_{\text{next}}, E_{\text{next}} \mid \mathcal{H}_t)$.

The distribution defined above describes when events occur (Andersen et al., 2012, Section II), e.g. that treatment with statins is applied at time t. In practice we also want to model what occurred in the event with finer granularity, e.g. the dosage of statins that is given. These features can be easily added on top of CEMs, and we describe this in section 4.3.

2.2 Problem Setting: Interventions on Latent Times

How should we formalize an intervention on time and estimate it using an autoregressive model? Consider the query "What if we treat a patient in two weeks?". Intuitively, we want to take the timing distribution of the treatment event $a \in \mathcal{E}$ and set it to a point mass in two weeks. To implement this using an autoregressive model, it is tempting to factorize the joint distribution over $(T_{\text{next}}, E_{\text{next}})$ as $P(T_{\text{next}} \mid E_{\text{next}}, \mathcal{H}_t) P(E_{\text{next}} \mid \mathcal{H}_t)$. This is called "time conditioned on mark" in Neural Temporal

Point Processes (Shchur et al., 2021). Under this factorization, a seemingly reasonable intervention is to replace $P(T_{\text{next}} \mid E_{\text{next}} = a, \mathcal{H}_t)$ with the point mass distribution on two weeks. A closer look at this option reveals that $P(E_{\text{next}} \mid \mathcal{H}_t)$, the distribution of the next event type, is invariant under this intervention. This is counterintuitive and problematic: changing treatment timing should affect the distribution over which event occurs next, e.g., death or hospital discharge. [2]

Instead, let us start from a foundational view of the problem. Following standard practice in causal modeling, we assign a variable and node in a Bayesian network to each intervenable object in the system. To this end T_e is a timing variable for each event $e \in \mathcal{E}$. An intervention on the "time of event a" is then an intervention on the mechanism that determines T_a. We model this with a conditional timing distribution $q_a(T_a \mid \mathcal{H}_t)$, and an intervention replaces it with a target distribution of interest. The variables T_a can be thought of as a (probabilistic) alarm clock that rings at times sampled from q_a, in which the event occurs and the clocks for all events $\mathcal{E}$ are reset. Hence, what affects these clocks are the past events, $\mathcal{H}_t$, and the values of T_e determine the type and time of the next event to occur. This is described by the graph in fig. 2.

Causal model via the competing events representation. Writing the distribution P induced by this graphical model while using the shorthand $\boldsymbol{T}_{\mathcal{E}} := [T_1, \ldots, T_{|\mathcal{E}|}]$ gives

$$P(E_{\text{next}}, \boldsymbol{T}_{\mathcal{E}} \mid \mathcal{H}_t) = q(E_{\text{next}} \mid \boldsymbol{T}_{\mathcal{E}}) \prod_{e \in \mathcal{E}} q_e(T_e \mid \mathcal{H}_t),$$

$$T_{\text{next}} = T_{E_{\text{next}}}. \tag{1}$$

Our goals are to learn this model from observed data, and use it to estimate trajectories under interventions. An interventional distribution $\tilde{P}$ replaces the conditional distribution $q_e(T_e \mid \mathcal{H}_t)$, with a distribution of interest $\tilde{q}_e(T_e \mid \mathcal{H}_t)$. We are then interested in estimation problems under these interventional distributions, e.g. the expected time to a myocardial infarction (MI) as defined below.

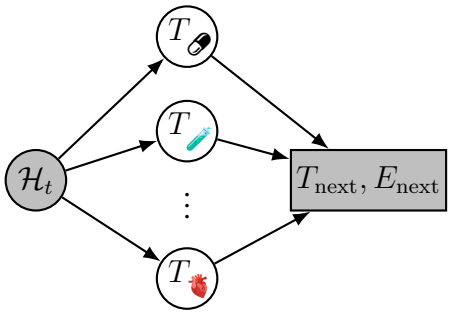
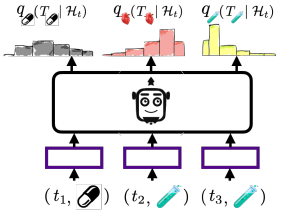


Figure 2: Graphical model depicting the time and type distribution of the next event, using latent times.

Definition 2.2 (interventions and estimands)**.** An intervention $\tilde{P}(\cdot \mid \mathcal{H}_t) = P(\cdot \mid do(\tilde{q}_a(T_a \mid \mathcal{H}_t)))$ on event $a \in \mathcal{E}$ in model P (see eq. (1)), replaces $q_a(t \mid \mathcal{H}_t)$ with a distribution $\tilde{q}_a(T_a \mid \mathcal{H}_t)$ while keeping all other components of P fixed. Let $f : \Omega \to \mathbb{R}$ be a measurable function of trajectories, we define the corresponding causal estimand $\mathbb{E}_{\tilde{P}}[f(\mathcal{H}) \mid \mathcal{H}_t]$.

3 Summary of the Method



(a)

Likelihood of $(t_{\text{next}}, e_{\text{next}})$

$$\mathcal{L}(\theta\,; t_{\text{next}}, e_{\text{next}}) = \log q_{e_{\text{next}}}(t_{\text{next}}; \theta) + \sum_{e \prec e_{\text{next}}} \log\left(1 - F_e[t_{\text{next}} + 1; \theta]\right) + \sum_{e \succ e_{\text{next}}} \log\left(1 - F_e[t_{\text{next}}; \theta]\right)$$

(b)

Sampling Next Event

Sample $t_e \sim q_e(T_e \mid \mathcal{H}_t; \boldsymbol{\theta}) \quad \forall e \in \mathcal{E}$

Set $e_{\text{next}} \leftarrow \arg\min_{\prec; e} t_e,\ t_{\text{next}} \leftarrow t_{e_{\text{next}}}$

(c)

Figure 3: (a) Example of a transformer producing distributions over latent times T_e. (b) The likelihood that we optimize to train a CEM, where for brevity, $\mathcal{H}_t$ is omitted and a discrete timeline is taken. (c) The next event sampling procedure from a trained CEM. The ordering $\prec$ is used to break ties between latent times when the minimum is not unique, and the likelihood is adjusted accordingly.

We present the implementation of CEMs here and then explain each component of the solution in section 4. The implementation follows eq. (1) by fitting the parameters $\boldsymbol{\theta}$ to learn distributions $\{q_e(T_e \mid \mathcal{H}_t; \boldsymbol{\theta})\}_{e \in \mathcal{E}}$. This specifies all the components of eq. (1) except $q(E_{\text{next}} \mid \boldsymbol{T}_{\mathcal{E}})$, which we fix to a point-mass distribution $q(E_{\text{next}} = e_{\text{next}} \mid \boldsymbol{T}_{\mathcal{E}} = \boldsymbol{t}_{\mathcal{E}}) = 1$ on the event with minimal time $e_{\text{next}} = \arg\min_{\prec; e} t_e$. Here, $\prec$ is an ordering of the event types $\mathcal{E}$, so $\arg\min_{\prec}$ chooses

[2] Similar challenges arise if we intervene on joint probabilities $P(T_{\text{next}}, E_{\text{next}} = a)$, since shifting time mass for one event type requires renormalizing the distribution, which may induce unintended changes in other event probabilities.

the event with minimum time from $\boldsymbol{t}_{\mathcal{E}}$, where the first minimizer according to $\prec$ is chosen if the minimum is non-unique. Training optimizes the log-likelihood, shown in fig. 3(b), where $F_e[t_e \mid \mathcal{H}_t; \theta] := q_e(T_e < t_e \mid \mathcal{H}_t)$ is the CDF of the distribution q_e. At inference time, latent times are sampled for each $e \in \mathcal{E}$ and the minimum is chosen (with ties broken according to $\prec$), see fig. 3(c). To estimate an effect, we replace the fitted q_a with $\tilde{q}_a$ of interest and sample complete trajectories $\{\mathcal{H}_i\}_{i=1}^m$, conditioned on $\mathcal{H}_t$, see fig. 4. The estimate is then $\mathbb{E}_{\tilde{P}}[f(\mathcal{H}) \mid \mathcal{H}_t] \approx m^{-1} \sum_i f(\mathcal{H}_i)$. Section 4 motivates and details each component of CEMs, Section 5 relates CEMs to intensity-based representations and prior work, and section 6 evaluates them.

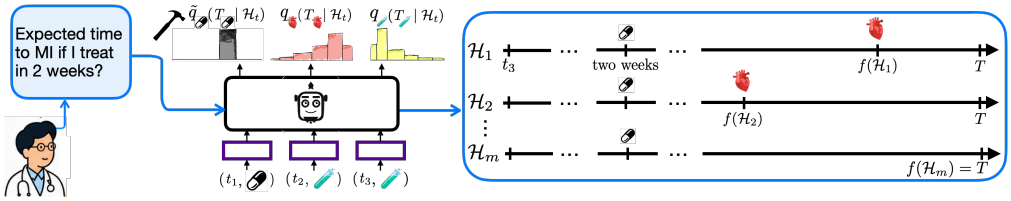


Figure 4: Example of an intervention on $q_{\text{💊}}$, and the trajectories sampled for effect estimation.

4 Competing Event Models

The latent times T_e and distributions $q_e(T_e \mid \mathcal{H}_t; \theta)$ are common in survival analysis and competing risks (Crowder, 2012; Kalbfleisch & Prentice, 2002; Tsiatis, 1975; Lee et al., 2018; Ranganath et al., 2016; Miscouridou et al., 2018), where another risk may cause an event before the risk of interest. CEMs extend these principles to sequence data and interventions on timing.

4.1 Why Events Compete

We begin by choosing a form for $q(E_{\text{next}} \mid \boldsymbol{T}_{\mathcal{E}})$ in eq. (1), based on assumed properties of interventions, defined by the following axioms.

Axiom 4.1. *For the causal model defined in eq. (1) with $\mathcal{T}$ as a finite set, let the CDFs of $q_e(T_e \mid \mathcal{H}_t)$ be denoted by $F_e[t_e \mid \mathcal{H}_t; \theta] := q_e(T_e < t_e \mid \mathcal{H}_t)$. Let $a \in \mathcal{E}$ and the corresponding interventional distributions of the form $\tilde{P}(\cdot \mid \mathcal{H}_t) = P(\cdot \mid do(\tilde{q}_a(T_a \mid \mathcal{H}_t)))$. For any choice of $\{q_e\}_{e \in \mathcal{E}}$ and an intervention $\tilde{q}_a$, consider the following axioms:*

a. (No backward effect) *If $\tilde{F}_a[s \mid \mathcal{H}_t] = F_a[s \mid \mathcal{H}_t]$ for all $t \leq s \leq t + \Delta$ where $\Delta > 0$, then $P(E_{\text{next}} = e, T_{\text{next}} = s) = \tilde{P}(E_{\text{next}} = e, T_{\text{next}} = s)$ for all $e \in \mathcal{E}$ and $t \leq s < t + \Delta$.*

b. (Expressivity) *Any distribution $P(E_{\text{next}}, T_{\text{next}} | \mathcal{H}_t)$ supported on $\mathcal{E} \times \mathcal{T} \cup (\textit{END}, \infty)$, can be realized by some $\{q_e\}_{e \in \mathcal{E}}$ where $q_e(\infty | \mathcal{H}_t) > 0$ for every $e \in \mathcal{E}$ with $P(E_{\text{next}} = e \mid \mathcal{H}_t) = 0$.*

Competing events as a consequence of axioms. Axiom 4.1(b) simply asks that our causal model can express any distribution over the observed events. For example, a choice that violates this is $q(E_{\text{next}} = e' \mid \boldsymbol{T}_{\mathcal{E}}) = 1$ for some specific $e' \in \mathcal{E}$, as it cannot express distributions where events other than e' occur. The axiom further asks that event types that do not occur can be represented by clocks that, with positive probability, never ring, which gives $T_e = \infty$ its intended meaning that e does not occur. Axiom 4.1(a) is also rather intuitive: intervening on what happens after a future time $t + \Delta$, should not affect what happens up until that time. [3] Lemma 4.1 that we give below and prove in appendix C shows that these two axioms characterize $q(E_{\text{next}} \mid \boldsymbol{T}_{\mathcal{E}})$ so it chooses an event with minimal time, which motivates CEMs as described in definition 4.1.

Lemma 4.1. *For the causal model in eq. (1) to satisfy axiom 4.1 for all settings of $\{q_e\}_{e \in \mathcal{E}}$ and interventions $\tilde{q}_a$ for any $a \in \mathcal{E}$, it must satisfy $q(E_{\text{next}} = e \mid \boldsymbol{T}_{\mathcal{E}} = \boldsymbol{t}_{\mathcal{E}}) = 0 \quad \forall e \notin \arg\min_{e'} t_{e'}$.*

Definition 4.1 (Competing Event Models)**.** A CEM is a parametric model of eq. (1), consisting of conditional mass (or density) functions, $\{q_e(T_e \mid \mathcal{H}_t; \theta)\}_{e \in \mathcal{E}}$, whose CDFs are denoted by $F_e[t_e \mid \mathcal{H}_t; \theta] := q_e(T_e < t_e \mid \mathcal{H}_t)$. The induced distribution $q(T_{\text{next}}, E_{\text{next}} \mid \mathcal{H}_t; \theta)$ is obtained by pushing forward $E_{\text{next}} \in \arg\min_{e \in \mathcal{E}} \{T_e\}$, $T_{\text{next}} = T_{E_{\text{next}}}$.

[3] In the continuous case where $\mathcal{T} = [0, T]$, this axiom is stated with equality for any measurable subset of $\mathcal{E} \times \mathcal{T}$.

CEMs in continuous time and no-concurrent-events. A canonical setting that is simple to model is when $\mathcal{T} = [0, \infty)$ and there are no concurrent events in the data. This is simple because models are based on absolutely continuous distributions q_e (e.g. mixtures of Gaussians) and then the probability of a tie $T_e = T_{e'}$ for a pair $e' \neq e$ is 0. So, the winning event $e_{\min} = \arg\min_e t_e$ is well-defined, and the expression we gave in lemma 4.1 simplifies further to $q(E_{\text{next}} = e_{\min} \mid \boldsymbol{T}_{\mathcal{E}} = \boldsymbol{t}_{\mathcal{E}}) = 1$. The log-likelihood of observing an event $(t_{\text{next}}, e_{\text{next}})$ is given by $\mathcal{L}(t_{\text{next}}, e_{\text{next}} \mid \mathcal{H}_t) = \log q_{e_{\text{next}}}(t_{\text{next}} \mid \mathcal{H}_t) + \sum_{e \neq e_{\text{next}}} \log\left(1 - F_e[t_{\text{next}} \mid \mathcal{H}_t]\right)$, which is composed of the likelihood that $T_{e_{\text{next}}}$ takes t_{next} and all other T_e take larger values. In appendix C.2 we also show there is a one-to-one correspondence between joint distributions $P(T_{\text{next}}, E_{\text{next}})$ and their competing event representations. Although models without concurrent events are useful, many datasets of interest, and particularly electronic health records, do have concurrent events. We turn to this case next.

4.2 CEMs with Concurrent Events and in Discrete Time

Concurrent events arise naturally when multiple lab measurements are taken at the same time, events occur at very close proximity and then recorded with the same timestamp, and many other cases. In this section, we wish to model such data in its natural form.

Atoms and discrete times. A simple and mathematically principled way to model concurrent events is to expand the original $\mathcal{E}$ to include subsets, resulting in $2^{|\mathcal{E}|}$ types. For $J = \{e'_1, \ldots, e'_L\}$, a CEM would have a distribution q_J of a latent time where the events J occur simultaneously. This puts us back in the no-concurrent-events setting in section 4.1, but it is intractable to implement: a distribution over $2^{|\mathcal{E}|}$ types is too large to store in memory even for a modest $|\mathcal{E}|$. The next simple solution is to represent simultaneous events autoregressively, as a sub-sequence of $|J|$ events, $(t_{\text{next}}, e'_1), (0, e'_2), \ldots, (0, e'_L)$. To generate such sequences, a model has to place a strictly positive probability on time 0, i.e., it is an *atom*. Introducing an atom avoids the exponential blowup of the naïve solution, but diverges from the no-concurrent-events setting. In models with finite $\mathcal{T}$ all time slots are atoms, and they are a natural setting to study them. Thus, in this section ***our focus is models where $\mathcal{T}$ is a finite set***, both because this is used in practice (e.g. bins $\mathcal{T} = \{[0, b_1], [b_1, b_2], \ldots\}$ are used in Renc et al. (2024) and in our experiments) and because it makes us handle atoms and concurrent events. To simplify notation, we take, without loss of generality, $\mathcal{T} = \{0, 1, \ldots, T\}$.

Tie-breaking, orderings and the likelihood. For a CEM, when $\{q_e\}_{e \in \mathcal{E}}$ distributions are over finite sets, they have atoms and there is a non-zero probability of ties in $\boldsymbol{T}_{\mathcal{E}}$. So unlike the no-concurrent-events case, $\arg\min_e T_e$ may be non-unique, and $q(E_{\text{next}} \mid \boldsymbol{T}_{\mathcal{E}})$ must break the ties between minimizers. For example, by spreading the probability equally between the winning events, or picking a type deterministically among the winners.

The choice of tie-breaking technique matters for the computational complexity of training an autoregressive model. Under such a model, the likelihood that events $J = \{e'_1, \ldots, e'_L\}$ occur simultaneously at time t_{next} is a sum of likelihoods over all permutations of these events, i.e. subsequences that represent events J. Writing this down explicitly, $\pi(L)$ is the set of permutations over L elements, we get the expression: $\sum_{\sigma \in \pi(L)} P(E_{\text{next}} = e'_{\sigma(1)}, T_{\text{next}} = t_{\text{next}} \mid \mathcal{H}_t) \prod_{j=2}^{L} P(E_{\text{next}} = e'_{\sigma(j)}, T_{\text{next}} = 0 \mid \mathcal{H}_t, (t_{\text{next}}, e'_{\sigma(1)}), \ldots, (0, e'_{\sigma(j-1)}))$. Computing this during training can be intractable due to the sum of $L!$ items. We consider a tie-breaking strategy that endows the model with an ordering $\prec$ over all types $\mathcal{E}$, where if $e \prec e'$ then e is always preferred in a tie with e'.

Definition 4.2. Let $\prec$ denote an ordering over the event types $\mathcal{E}$ and $\min_{\prec}$ the operator that chooses a minimum element while breaking ties using $\prec$. A CEM under the ordering $\prec$ is the model in eq. (1) where $q(E_{\text{next}} = e \mid \boldsymbol{T}_{\mathcal{E}} = \boldsymbol{t}_{\mathcal{E}}) = \mathbf{1}[e = \arg\min_{\prec, e'} t_{e'}]$.

An attractive property of orderings is that they let us drop the sum over all permutations $\pi(L)$ and assuming that $e'_1 \prec e'_2 \prec \ldots \prec e'_L$, the new expression for the likelihood above reduces to a single summand $P(E_{\text{next}} = e'_1, T_{\text{next}} = t_{\text{next}} \mid \mathcal{H}_t) \prod_{j=2}^{L} P(E_{\text{next}} = e'_j, T_{\text{next}} = 0 \mid \mathcal{H}_t, (t_{\text{next}}, e'_1), \ldots, (0, e'_{j-1}))$. The likelihood for each event in a trajectory (regardless of whether it is concurrent with other events as part of J) under an ordering $\prec$ is then given by

$$\mathcal{L}(t_{\text{next}}, e_{\text{next}} \mid \mathcal{H}_t) = \log q_{e_{\text{next}}}(t_{\text{next}} \mid \mathcal{H}_t) + \sum_{e \prec e_{\text{next}}} \log\left(1 - F_e[t_{\text{next}} + 1 \mid \mathcal{H}_t]\right) + \sum_{e \succ e_{\text{next}}} \log\left(1 - F_e[t_{\text{next}} \mid \mathcal{H}_t]\right). \tag{2}$$

Interventions in the presence of concurrent events. Beyond computation, tie-breaking dictates the semantics of q_a and interventions on it. In appendix C.3 we discuss this extensively, and here we aim to give an intuitive summary. Under $\prec$, $q_a(t_a \mid \mathcal{H}_t)$ is the probability that a occurs at t_a conditioned on no events occurring in (t, t_a) *and also no preceding events* $e \prec a$ *occurring at* t_a. Consequently, a's position in $\prec$ determines what concurrent information a decision on when to do a can leverage. The two intuitive examples being:

- **Placing** a **last in** $\prec$: $q_a(t_a \mid \mathcal{H}_t)$ is the probability of a at time t_a, and that after it in the sequence there are no other events generated at time t_a.
- **Placing** a **first in** $\prec$: $q_a(t_a \mid \mathcal{H}_t)$ is the probability of a at time t_a, and it is unknown whether other events $e \succ a$ occur simultaneously. In other words, it is the probability that any $J \subseteq \mathcal{E}$ where $a \in J$ occurs at time t_a.

In our statins and MI example, say we take 🧪 $\prec$ 💊. This means an intervention on 💊 can express interventions such as prescribing statins conditionally on a high cholesterol reading observed at the exact same timestamp t (via $\tilde{q}_{💊}(0 \mid \mathcal{H}_t, (t_{\text{next}}, 🧪)) = 1$). On the other hand, an intervention on 🧪 can set the lab test at some time t_{next} and have the prescriptions or other labs that usually occur with it follow suit. Both interventions are useful in practice, but in our experiments, we focus on interventions where the target event is last in $\prec$.

Crucially, ordering does not restrict observational expressivity of the model. As we state below, any choice of $\prec$ can represent any joint distribution $P(T_{\text{next}}, E_{\text{next}})$, it simply defines the factorization and interventional semantics. The result is a direct consequence of lemma C.7 in the appendix, where we give a procedure to transform a joint distribution to a CEM under any chosen $\prec$.

Lemma 4.2. *Let* $P(T_{\text{next}}, E_{\text{next}})$ *be a joint distribution. There exists a CEM under ordering* $\prec$, $\{q_e\}_{e \in \mathcal{E}}$, *that represents it. The representation is unique for all values* $q_e(t)$ *where* $P(T_{\text{next}} = t, E_{\text{next}} = e) > 0$, *and the CEM satisfies axiom 4.1.*

4.3 Implementation and Scalability

Several details require attention when implementing CEMs at scale: adding event features, efficient gradient computation in training and computing effect estimates at inference time.

Training CEMs at scale. Denoting $|\mathcal{T}|$ as the number of parameters required to describe the distribution q_e, we see that to implement a CEM and optimize $\mathcal{L}(\mathcal{H}; \theta)$, a neural network has to output the parameters of $\{q_e\}_{e \in \mathcal{E}}$, a tensor of size $b \times n \times |\mathcal{E}| \times |\overline{\mathcal{T}}|$ where n is the length of the sequence and b is the batch size. When $|\mathcal{E}|$ is large, for example, when we train a CEM on $|\mathcal{E}| > 80,000$ and $n > 100$ in section 6, the memory of a single GPU would be well exhausted. In this scenario, the objective in eq. (2) is attractive because it decomposes over $\mathcal{E}$, so that during optimization we can obtain an unbiased approximation of the gradient by only computing the logits for a randomly drawn subset of $\mathcal{E}$. Specifically, in our implementation we optimize the likelihood of observed events $q_{e_j}(t_j \mid \mathcal{H}_{t_{j-1}})$ in full, while subsampling $K < |\mathcal{E}|$ codes for the log-survival terms $\log(1 - F_e[t_j \mid \mathcal{H}_{t_{j-1}}])$ of eq. (2) and rescaling this part of the loss by $|\mathcal{E}|/K$. Note that this strategy can be applied due to the competing events representation, since likelihoods for other representations such as the logits of the raw joint distribution $P(T_{\text{next}}, E_{\text{next}})$ do not decompose similarly. Lastly, our experiments use a discretized timeline It is also worth noting that while our experiments use a discretized timeline where it is striaghtforward to calculate the CDF F_e. But continuous distributions such as mixture densities (Han et al., 2022) can also be used without implementing integration, via an unbiased gradient estimator (appendix C).

Interventions at inference time and supporting multiple orderings. To apply an intervention on a fitted model $q(\cdot; \theta)$, we can manually set the time T_a to a value $t + \delta$ and follow the min-time sampling procedure to estimate queries of the form "what if we treat at time $t + \delta$?". More generally, replacing $q_a(\cdot; \theta)$ with a closed-form alternative, as we do in our experiments, is straightforward.

Our implementation trains a shared transformer representation and a separate linear prediction head θ_e for each $e \in \mathcal{E}$. An important architectural consideration is how to adjust our model to an ordering $\prec$ chosen at inference time. To this end we specify two methods. ***(i) Architecture based.*** The method we use in section 6 draws random permutations on $\mathcal{E}$ during training, embeds the corresponding ordering $\prec$ as prefix tokens at the start of each trajectory using cross-attention with a small set of

learned query vectors. During pretraining we optimize the likelihood in eq. (2) according to these drawn orderings and learn a model that handles arbitrary orderings at inference time. Figure 6 in appendix D describes this architecture. ***(ii) Representation conversion.*** Appendix C gives an $O(|\mathcal{T}| \times |\mathcal{E}|)$ procedure that converts $P(T_{\text{next}}, E_{\text{next}}; \theta)$ to a competing events representation under a desired $\prec$, trading a simpler model for computation at inference time. Our experiments show that the architectural solution is stable for the problem we study.

Additional features and interventions on what to do. As pointed out in the introduction, it is often desired to intervene on *what* to do and not just when to do it. That is, intervening on features of the event, such as the dosage of a drug. To this end we include additional features $z_j \in \mathbb{R}$ for each event (t_j, e_j), so the trajectories are composed of triplets $\{(t_j, e_j, z_j)\}_{j=1}^n$. Then each event with a feature z_j is encoded by 2 items in the sequence, the first represents (t_j, e_j), and the second encodes z_j. That is, the prediction of z_j aims to simulate the distribution $P(Z_{\text{next}}|\mathcal{H}_t, t_{\text{next}}, e_{\text{next}})$.

To conclude, algorithm 2 in appendix D gives the full pseudo-code for training, sampling and effect estimation, including the ordering from section 4.2 and additional numerical features.

4.4 Identifiability of Causal Estimands

In most real-world settings we do not observe all events relevant to an estimation problem; for example, treatments may depend on a patient's appearance, which is not recorded. We therefore extend the observed types $\mathcal{E}$ with unobserved types $\mathcal{U}$, whose trajectories we denote by $\mathcal{H}^{\mathcal{U}}$, so that clocks may depend on them, $q_e(T_e \mid \mathcal{H}_t, \mathcal{H}_t^{\mathcal{U}})$, while the CEM we fit learns the marginal distribution over $\mathcal{E}$. The events of type e in a trajectory $\mathcal{H}$ are denoted $\mathcal{H}^e$.

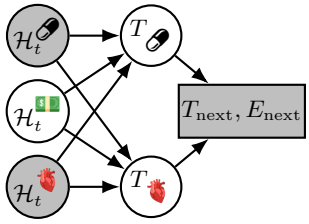


Figure 5: Example of unobserved confounding that violates the identifiability conditions in assumption 4.1. Gray nodes are observed, white nodes are not.

Since the CEM learns a marginal distribution, it may produce an incorrect estimate of the causal effect. Here we specify conditions under which CEMs do recover the correct effect. These were developed in Røysland et al. (2022); Didelez (2008) and are essentially a generalization of no-unobserved-confounding, or more precisely of sequential exchangeability (Hernán & Robins, 2010) to timing effects. Unobserved confounders are unobserved events whose history affects the future times of both the treatment and the outcome. Looking back at our example of statin treatment to prevent MI, these are events that affect the timing of both the outcome $y =$ and $a =$. Let $f(\mathcal{H}) = f(\mathcal{H}^y)$ denote the time an MI occurs in trajectory $\mathcal{H}$, and consider an intervention on a (as in definition 2.2).[4] One example of such a confounding event is a patient's past salary, $u =$. It affects a patient's financial means to obtain medication, and also affects risk of MI through lifestyle factors such as better nutrition. This is described in fig. 5. Intuitively, if individuals with high salaries are both healthier due to their lifestyle and tend to take the drug more often, then estimation via a model that does not take past salaries into account may attribute the improved outcomes to the drug even though in reality it has a non-significant effect. The assumption below restricts the regimes in which we guarantee correct effect estimation to those where such unobserved confounding is absent. For an unobserved confounder u, condition *(i)* roughly says that u holds no information about the treatment a, while condition *(ii)* says u does not hold information about the outcome.[5] We additionally assume overlap, $P \gg \tilde{P}$, i.e., that $\tilde{P}$ is absolutely continuous w.r.t P.

Assumption 4.1. *Let $f(\mathcal{H}) = f(\mathcal{H}^y)$ where $y \in \mathcal{E}$ is an outcome type. We assume the unobserved processes can be decomposed as $\mathcal{U} = [u_1, \ldots, u_l]$ such that for each u_i, where we denote $\mathcal{U}_{>i} = [u_{i+1}, \ldots u_l]$, it holds either that* (i) $q_a(T_a \mid \mathcal{H}_t, \mathcal{H}_t^{u_i}, \mathcal{H}_t^{\mathcal{U}_{>i}}) = q_a(T_a \mid \mathcal{H}_t, \mathcal{H}_t^{\mathcal{U}_{>i}})$, *or* (ii) $q_e(T_e \mid \mathcal{H}_t, \mathcal{H}_t^{u_i}, \mathcal{H}_t^{\mathcal{U}_{>i}}) = q_e(T_e \mid \mathcal{H}_t, \mathcal{H}_t^{\mathcal{U}_{>i}})$ *for all* $e \in \{\mathcal{E} \setminus a\} \cup \mathcal{U}_{>i}$.

Assumption 4.1 fails in fig. 5: With $\mathcal{U} = \{$ $\}$, (i) requires removing the edge $\mathcal{H}_t^{} \to T_{}$, and (ii) requires removing $\mathcal{H}_t^{} \to T_{}$. The ordered traversal of $\mathcal{U}$ extends this to confounding through

[4] Formally, with unobserved events $\mathcal{U}$, definition 2.2 should be modified so an intervention replaces $q_a(T_a \mid \mathcal{H}_t, \mathcal{H}_t^u)$ and not $q_a(T_a \mid \mathcal{H}_t)$.

[5] Condition *(ii)* is stated on $\mathcal{E} \setminus a$ and not just y, since under sequential effects, the graph is unrolled and effect of u on y can flow from other sources. Assumptions based on sequential exchangeability (Hernán & Robins, 2010) can also be phrased.

sequences of unobserved variables. Unobserved confounding also makes the T_e dependent given $\mathcal{H}_t$, so the independent latent times (conditionally on $\mathcal{H}_t$) that CEMs learn may be an incorrect world model. In Pearl's terminology, it fits a Markovian model (Pearl, 2009, Sec. 1.4.2), while in reality the data generating process it unlikely to be Markovian w.r.t the observed data. This cannot be verified from observed data since a Markovian model can express any $P(T_{\text{next}}, E_{\text{next}})$, see Tsiatis (1975) in the context of latent times. In appendix C.5 we prove that under assumption 4.1 the CEM still gives a correct estimate of $\mathbb{E}_{\tilde{P}}[f(\mathcal{H}) \mid \mathcal{H}_t]$. Its clocks should still be interpreted with care: For example, we should not assign the distribution $q_{❤}(t_{❤} \mid \mathcal{H}_t; \theta)$ a causal interpretation as the probability density that MI occurs at $t_{❤}$ when all other events are *prevented*. This is because this intervention acts on all $\mathcal{E} \setminus \{❤\}$ at once, and it is unlikely that none of them are confounded with MI. Rather, it is the probability that MI occurs at t given that no other event occurs before it.

5 Connection to Intensity Based Representation and Related Work

5.1 Interventions on Intensities vs. Interventions on Competing Clocks

The intensity $\lambda_e(t)$ is the probability that e occurs at t, conditioned on no event occurring beforehand; in discrete time, $\lambda_e(t_{\text{next}}) = P(t_{\text{next}}, e)/P(T_{\text{next}} \geq t_{\text{next}})$. Several works define interventions on time by replacing $\lambda_e \to \tilde{\lambda}_e$ (Lok, 2008; Didelez, 2008; Rytgaard et al., 2022; Røysland, 2011; Qu et al., 2023; Schulam & Saria, 2017; Wald et al., 2025). We would expect this formalism and the CEM one to coincide: : if $\tilde{P}$ results from $q_a \to \tilde{q}_a$, then $\lambda_e = \tilde{\lambda}_e$ for all $e \neq a$.

It is important to understand how the two representations relate and what to consider when implementing a model for one or the other.

This holds in continuous time without concurrent events (lemma C.2), but in discrete time only when a is last under $\prec$. However, for discretized time it is no longer true. The correspondence only holds when a is last under the ordering $\prec$. Furthermore, when there are concurrent events the interventions on intensities become unintuitive. We provide intuition and technical details in section C.3.2, as well as a proof for the following result which states that given a choice of ordering, any intervention in an intensity-based representation can be expressed as a competing-events based intervention.[6]

Theorem 5.1. *Let λ and $\tilde{\lambda}$ be intensities that represent distributions over the next time and event, before and after an intervention on λ_a. There exists competing event models $q, \tilde{q}$ and ordering $\prec$ under which the models induce the same distributions as $\lambda, \tilde{\lambda}$ respectively, and $q_e = \tilde{q}_e \; \forall e \neq a$.*

Implementation considerations. In practice, sampling from intensity-based models in continuous time requires algorithms such as thinning (Lewis & Shedler, 1979), which are non-standard in autoregressive models, and their normalization does not allow the unbiased subsampling of event types that we use for CEMs (see section 4.3).

5.2 Related Work

Neural Temporal Point Processes and Predicting Next Times. Many generative models for irregular time series neither implement interventions nor handle concurrent events. Neural Hawkes processes (Mei & Eisner, 2017) and recent foundation models (Shmatko et al., 2025; Steinberg et al., 2023) parameterize intensities λ_e directly, requiring non-standard sampling such as thinning (Lewis & Shedler, 1979) and assuming continuous time where concurrent events have probability zero. Intensity-free models (Shchur et al., 2019) learn $P(T_{\text{next}} \mid \mathcal{H}_t; \theta)$ instead, which is common in recent transformer-based healthcare models (McDermott et al., 2023; Song et al., 2024; Yang et al., 2023; Labach et al., 2023; Hill et al., 2023; Pang et al., 2024; Renc et al., 2024). However, as discussed in section 2, this governs the next time for an event of *any* type, making it *unclear how to implement an intervention* on treatment timing.

Interventions on Time. The formalism of interventions on intensities of events was developed in (Lok, 2008; Didelez, 2008), with subsequent statistical tools including targeted estimators (Rytgaard et al., 2022; Nie et al., 2021) and neural approaches (Vanderschueren et al., 2023; Hess & Feuerriegel, 2025). Most neural solutions condition on history $\mathcal{H}_{t_0}$ up to some time t_0 and specify a static

[6] No single $q(E_{\text{next}} \mid \boldsymbol{T}_{\mathcal{E}})$ gives this correspondence for between interventions on q_a and λ_a for all $a \in \mathcal{E}$, see lemma C.4.

intervention for all future treatments (e.g., what if we treat in two months, regardless of what happens in the interim). By contrast, the interventions we consider are *dynamic*: they depend on $\mathcal{H}_t$ up to the treatment time (e.g., treating sooner if the condition worsens) and are often called *policies*.

Off-policy Evaluation and Model Based RL. Effect estimation of policies from observational data is studied as Off-Policy Evaluation (OPE) (Uehara et al., 2025; Oberst & Sontag, 2019), predominantly on regularly-spaced timelines. Interventions on *when* to act that arise naturally in the irregularly sampled data we consider, require time to be part of the policy. Some RL work explores this in online settings (Upadhyay et al., 2018; Qu et al., 2023). In the off-policy setting, the closest work is Wald et al. (2025), who fit a value function for a specific policy and outcome. CEMs instead provide a world model, enabling evaluation of multiple policies and outcomes without retraining.

6 Experiments

Evaluating causal estimators requires simulators to provide ground truth interventional outcomes. To this end, we work with a commonly used tumor growth simulator (Bica et al., 2020; Seedat et al., 2022; Vanderschueren et al., 2023; Hess & Feuerriegel, 2025) to do off-policy evaluation. CEMs are also distribution estimators, and the correctness of any causal estimate obtained with them relies on how well they learn the data distribution. This property can be evaluated on real-world observational data; therefore we use risk prediction tasks on MIMIC-IV.

6.1 Datasets and Methods

The cancer simulator has $N = 10{,}000$ trajectories and $|\mathcal{E}| = 4$ (OPE task), and MIMIC-IV has $N = 246{,}773$ and $|\mathcal{E}| = 83{,}065$ (risk prediction). We use the MEDS standard and MEDS-torch package (Arnrich et al., 2024; Oufattole et al., 2024), a suite of tools to process event streams of the form (t, e, z). Performance of baseline methods on the MIMIC-IV tasks is also obtained from the package. ***(i) Tumor volume simulator.*** The simulated dataset comes from a pharmacokinetic model of tumor growth in cancer (Geng et al., 2017; Bica et al., 2020). This setting was adapted in prior works to study effect estimation with irregularly sampled times, yet with non-dynamic policies. We describe the task for this experiment and give the details of the simulation environment in appendix D. The simulator works in discrete time $t \in [T]$ where $T = 20$, and irregular event times are induced by the tumor volume being unobserved at most times. We have $\mathcal{E} = \{\mathrm{x}, \mathrm{a}_{\text{chemo}}, \mathrm{a}_{\text{radio}}\}$ and another static feature corresponding to the type of patient, where the mark $z_t \in \mathbb{R}_+$ for x is the tumor volume. The volume is observed with probability $\sigma((\bar{z}_{t-d:t-1}/d_{\max}) - 1.5)$, where $\bar{z}_{t-d:t-1}$ is the average (potentially unobserved) tumor volume over the last d time steps, and $d_{\max}$ is the maximum considered volume. The treatments are radiotherapy, chemotherapy, or combined therapy when the two events are concurrent. Treatment policies are determined by two parameters (γ, β), where each type of treatment is assigned with probability $\sigma(\gamma(x_{\text{last}} - \beta) + t - t_{\text{last}})$. Here, x_{last} is the last observed volume and t_{last} is the time of the last treatment. β controls how often treatments are applied, while γ controls the dependence of treatment assignment on tumor volume. **The task** is to predict tumor volume at time T given the history $\mathcal{H}_t$, if for times greater than t, the patient will be treated with a policy of interest $(\gamma, \beta)_{\text{int}}$. ***(ii) MIMIC-IV*** (Johnson et al., 2020)**.** We extract the dataset into MEDS format, which contains trajectories of medical event codes (labs, prescriptions, admissions etc.), their timestamps, and associated numerical values. Each trajectory corresponds to a patient. Following Oufattole et al. (2024), we filter trajectories with fewer than 10 events and remove codes that appear in fewer than $10,000$ patients. We discretize time differences between events into 74 bins based on quantiles. CEMs are pretrained with next-event prediction on these trajectories. **The tasks** we evaluate are medical risk prediction tasks, including in-hospital, ICU, and post-discharge mortality prediction, as well as 30-day hospital readmission.

Architecture. We implement CEMs by modifying a decoder-only transformer based on GPT-2. Token dimensions and number of layers vary between the two datasets, but we use the same backbone for all methods within the same dataset. For CEMs we make the modifications described in section 4.3: $|\mathcal{E}|$ prediction heads are introduced for the latent times T_e; each head is a rank-8 matrix multiplication of the last-layer representation. In MIMIC-IV, we subsample event types from $\mathcal{E}$ during optimization with $K = 1024$. We provide additional details about the implementation in appendix E.

Baselines and metrics. *Cancer simulator.* We compare CEMs to 3 methods: Empirical Risk Minimization / Monte-Carlo (ERM/MC) predicts the final volume from each prefix in the series, ignoring the intervention on the treatment policy used at evaluation; FQE (Le et al., 2019) performs

Table 1: ***(a)*** NRMSE for tumor growth simulation. *Top table:* CEMs, EDQ and a direct forecasting model are comparable under the observational policy they are trained on. Under an intervention, CEM is more stable than the baselines. *Bottom table:* Estimation for CEMs is inaccurate when intervening incorrectly, or encoding a wrong order for tie breaking. ***(b)*** AUROC for medical risk prediction tasks, comparing finetuning under pretraining methods and a supervised model.

$(\gamma, \beta)_{\text{obs}}$	ERM / MC	FQE	EDQ	CEM
	$(\gamma, \beta)_{\text{int}} = (6, 0.5)$			
$(6, 0.5)$	**0.180** ± 0.017	0.560 ± 0.021	**0.204** ± 0.013	**0.189** ± 0.009
$(10, 0.75)$	0.327 ± 0.015	0.707 ± 0.017	0.262 ± 0.017	**0.184** ± 0.011
	CEM ablations			
	Wrong policy $(6, 0.75)$		Wrong tie break	
$(10, 0.75)$	0.228 ± 0.010		0.584 ± 0.015	

(a)

Method	Hosp	ICU	Discharge	Readmission
OCP	87.3 ± 0.4	75.8 ± 0.4	85.6 ± 0.1	71.5 ± 0.2
EBCL	86.5 ± 0.2	77.4 ± 0.7	85.6 ± 0.1	71.8 ± 0.1
AR	74.0 ± 0.6	65.1 ± 2.2	76.0 ± 1.8	64.4 ± 0.5
Value	**88.2** ± 0.1	77.3 ± 0.8	86.2 ± 0.1	72.1 ± 0.1
Super.	87.1 ± 0.9	77.4 ± 1.1	85.4 ± 0.2	71.5 ± 0.2
CEM	86.6 ± 0.1	**79.4** ± 0.3	**86.6** ± 0.1	**73.1** ± 0.1

(b)

dynamic programming in discrete time; EDQ (Wald et al., 2025) is a similar method but designed to handle interventions on time and irregularly sampled events. Appendix D describes the methods in more detail. For evaluation, we set fixed prediction times $t_1, \ldots, t_k$ in $[0, T]$, take each trajectory from the test set, and let the models predict the volume at time T from the history $\mathcal{H}_{t_i}$ for $i \in [k]$. For CEMs, prediction is made by generating 10 trajectories and averaging the volume estimate they produce at time T. We summarize the errors by taking the normalized RMSE over all test trajectories and time steps. *MIMIC-IV.* We compare finetuning a pretrained CEM against several pretraining baselines. These include two contrastive learning approaches, OCP (Jeong et al., 2023) and EBCL (Agrawal et al., 2022), as well as two forecasting based methods: Value forecasting (Tipirneni & Reddy, 2022) and auto-regression (AR).[7] We also include a supervised model trained on the same architecture, implemented and evaluated by Oufattole et al. (2024). We use the MEDS-torch repository to train our model with the same backbone and pipeline as the baselines. In supervised tasks, CEMs are fine-tuned with a single prediction head identical to the baselines.

6.2 Results

Table 1(a) reports the results on the tumor growth simulator. The first row of the top table evaluates models trained on data generated with the policy $(\gamma, \beta)_{\text{obs}} = (6, 0.5)$ and tested on trajectories generated with the same policy. In this case, $(\gamma, \beta)_{\text{obs}} = (\gamma, \beta)_{\text{int}}$, i.e., $P = \tilde{P}$. The methods achieve similar errors indicating comparable predictive performance under their training distribution, except for FQE where gradients do not propagate the errors efficiently from the label at the end of the trajectory to set predictions in the beginning. In the second row, we keep $(\gamma, \beta)_{\text{int}}$ fixed but train the model on trajectories generated under a different behavior policy $(\gamma, \beta)_{\text{obs}} = (10, 0.75)$. FQE and EDQ are provided with the policy $(\gamma, \beta)_{\text{int}}$ during training and account for the shift via dynamic programming. CEM receives the new policy at test-time and generates data using it, whereas ERM/MC directly outputs its prediction without accounting for the policy shift. As expected, CEMs achieve a similar error to the models trained on data from $(\gamma, \beta)_{\text{int}}$, while ERM/MC sees a considerably higher error. FQE underperforms again for similar reasons, while EDQ achieves the second-best performance. The bottom table shows two ablations: (i) plugging an incorrect policy into the CEM (changing γ but not β), and (ii) mishandling tied events by randomizing the order of tokens for tied events during training without encoding the tie-breaking order. Both ablations increase error relative to the correctly specified model. Appendix D has additional results for CEMs on other policies and a breakdown of prediction errors as a function of the prediction time t_k.

Table 1(b) presents finetuning results on MIMIC-IV. Value forecasting performs best for in-hospital mortality prediction, while CEMs achieve the highest AUROC on the remaining tasks. These results demonstrate that CEMs provide an effective parameterization for modeling interventions while learning representations useful for downstream tasks.

Conclusion. CEMs are world models for irregularly timed events, in which interventions on when and what to do amount to replacing a single clock. CEMs achieve the highest AUROC on 3 of 4 MIMIC-IV tasks, showing they also learn representations useful for downstream tasks. Limitations, including the identifiability assumptions and the need for validation on real-world effect estimation questions, are discussed in appendix A.

[7] Value forecasting encodes one time window and predicts the codes and values in the next. AR predicts the next event's code, value, and time given the prefix, with separate prediction heads.

Acknowledgments

The authors wish to thank Aahlad Puli and Shalmali Joshi for discussions that helped shape the final version of the paper. We also thank Carmel Pe'er, Raghav Singhal, Mark Goldstein, and Shauli Ravfogel for useful discussions.

Dr. Yoav Wald is the Ravitz Foundation Faculty Fellow. This work was partly supported by the NIH/NHLBI Award R01HL148248, NSF Award 1922658 NRT-HDR: FUTURE Foundations, Translation, and Responsibility for Data Science, NSF CAREER Award 2145542, ONR N00014-23-1-2634, NIH R01CA296388, NSF 2404476, Optum, the Gates Foundation, and Apple. This work was also supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) with a grant funded by the Ministry of Science and ICT (MSIT) of the Republic of Korea in connection with the Global AI Frontier Lab International Collaborative Research Agreement. (No. RS-2024-00469482 & RS-2024-00509279).

References

Monica N Agrawal, Hunter Lang, Michael Offin, Lior Gazit, and David Sontag. Leveraging time irreversibility with order-contrastive pre-training. In *International Conference on Artificial Intelligence and Statistics*, pp. 2330–2353. PMLR, 2022.

Per K Andersen, Ornulf Borgan, Richard D Gill, and Niels Keiding. *Statistical models based on counting processes*. Springer Science & Business Media, 2012.

Bert Arnrich, Edward Choi, Jason Alan Fries, Matthew BA McDermott, Jungwoo Oh, Tom Pollard, Nigam Shah, Ethan Steinberg, Michael Wornow, and Robin van de Water. Medical event data standard (meds): Facilitating machine learning for health. In *ICLR 2024 Workshop on Learning from Time Series For Health*, pp. 03–08, 2024.

Ioana Bica, Ahmed M Alaa, James Jordon, and Mihaela van der Schaar. Estimating counterfactual treatment outcomes over time through adversarially balanced representations. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=BJg866NFvB.

Irene Y Chen, Shalmali Joshi, Marzyeh Ghassemi, and Rajesh Ranganath. Probabilistic machine learning for healthcare. *Annual review of biomedical data science*, 4(1):393–415, 2021.

Martin J Crowder. *Multivariate survival analysis and competing risks*. CRC Press, 2012.

Daryl J Daley and David Vere-Jones. *An introduction to the theory of point processes: volume I: elementary theory and methods*. Springer, 2003.

Vanessa Didelez. Graphical models for marked point processes based on local independence. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 70(1):245–264, 2008.

Changran Geng, Harald Paganetti, and Clemens Grassberger. Prediction of treatment response for combined chemo-and radiation therapy for non-small cell lung cancer patients using a biomathematical model. *Scientific reports*, 7(1):13542, 2017.

Xintian Han, Mark Goldstein, and Rajesh Ranganath. Survival mixture density networks. In *Machine Learning for Healthcare Conference*, pp. 224–248. PMLR, 2022.

Miguel A Hernán and James M Robins. *Causal inference*. CRC Boca Raton, FL, 2010.

Konstantin Hess and Stefan Feuerriegel. Stabilized neural prediction of potential outcomes in continuous time. In *The Thirteenth International Conference on Learning Representations*, 2025.

Brian L Hill, Melikasadat Emami, Vijay S Nori, Aldo Cordova-Palomera, Robert E Tillman, and Eran Halperin. Chiron: A generative foundation model for structured sequential medical data. In *Deep Generative Models for Health Workshop NeurIPS 2023*, 2023.

Hyewon Jeong, Nassim Oufattole, Matthew Mcdermott, Aparna Balagopalan, Bryan Jangeesingh, Marzyeh Ghassemi, and Collin Stultz. Event-based contrastive learning for medical time series. *arXiv preprint arXiv:2312.10308*, 2023.

Alistair Johnson, Lucas Bulgarelli, Tom Pollard, Steven Horng, Leo Anthony Celi, and Roger Mark. Mimic-iv. *PhysioNet. Available online at: https://physionet. org/content/mimiciv/1.0/(accessed August 23, 2021)*, pp. 49–55, 2020.

Shalmali Joshi, Iñigo Urteaga, Wouter AC van Amsterdam, George Hripcsak, Pierre Elias, Benjamin Recht, Noémie Elhadad, James Fackler, Mark P Sendak, Jenna Wiens, et al. Ai as an intervention: improving clinical outcomes relies on a causal approach to ai development and validation. *Journal of the American Medical Informatics Association*, pp. ocae301, 2025.

John D Kalbfleisch and Ross L Prentice. *The statistical analysis of failure time data*. John Wiley & Sons, 2002.

Alex Labach, Aslesha Pokhrel, Xiao Shi Huang, Saba Zuberi, Seung Eun Yi, Maksims Volkovs, Tomi Poutanen, and Rahul G Krishnan. Duett: dual event time transformer for electronic health records. In *Machine Learning for Healthcare Conference*, pp. 403–422. PMLR, 2023.

Hoang Le, Cameron Voloshin, and Yisong Yue. Batch policy learning under constraints. In *International Conference on Machine Learning*, pp. 3703–3712. PMLR, 2019.

Changhee Lee, William Zame, Jinsung Yoon, and Mihaela Van Der Schaar. Deephit: A deep learning approach to survival analysis with competing risks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

PA W Lewis and Gerald S Shedler. Simulation of nonhomogeneous poisson processes by thinning. *Naval research logistics quarterly*, 26(3):403–413, 1979.

Rui Li, Stephanie Hu, Mingyu Lu, Yuria Utsumi, Prithwish Chakraborty, Daby M Sow, Piyush Madan, Jun Li, Mohamed Ghalwash, Zach Shahn, et al. G-net: a recurrent network approach to g-computation for counterfactual prediction under a dynamic treatment regime. In *Machine Learning for Health*, pp. 282–299. PMLR, 2021.

Judith J Lok. Statistical modeling of causal effects in continuous time. *The Annals of Statistics*, pp. 1464–1507, 2008.

Matthew McDermott, Bret Nestor, Peniel Argaw, and Isaac S Kohane. Event stream gpt: a data pre-processing and modeling library for generative, pre-trained transformers over continuous-time sequences of complex events. *Advances in Neural Information Processing Systems*, 36: 24322–24334, 2023.

Hongyuan Mei and Jason M Eisner. The neural hawkes process: A neurally self-modulating multivariate point process. *Advances in neural information processing systems*, 30, 2017.

Valentyn Melnychuk, Dennis Frauen, and Stefan Feuerriegel. Causal transformer for estimating counterfactual outcomes. In *International Conference on Machine Learning*, pp. 15293–15329. PMLR, 2022.

Xenia Miscouridou, Adler Perotte, Noémie Elhadad, and Rajesh Ranganath. Deep survival analysis: Nonparametrics and missingness. In *Machine Learning for Healthcare Conference*, pp. 244–256. PMLR, 2018.

Xinkun Nie, Emma Brunskill, and Stefan Wager. Learning when-to-treat policies. *Journal of the American Statistical Association*, 116(533):392–409, 2021.

Michael Oberst and David Sontag. Counterfactual off-policy evaluation with gumbel-max structural causal models. In *International Conference on Machine Learning*, pp. 4881–4890. PMLR, 2019.

Nassim Oufattole, Teya Bergamaschi, Pawel Renc, Aleksia Kolo, Matthew B.A. McDermott, and Collin Stultz. MEDS-torch: An ML pipeline for inductive experiments for EHR medical foundation models. In *NeurIPS Workshop on Time Series in the Age of Large Models*, 2024. URL https://openreview.net/forum?id=uwnK2LZtjX.

Chao Pang, Xinzhuo Jiang, Nishanth Parameshwar Pavinkurve, Krishna S Kalluri, Elise L Minto, Jason Patterson, Linying Zhang, George Hripcsak, Gamze Gürsoy, Noémie Elhadad, et al. Cehr-gpt: Generating electronic health records with chronological patient timelines. *arXiv preprint arXiv:2402.04400*, 2024.

Judea Pearl. *Causality*. Cambridge university press, 2009.

Chao Qu, Xiaoyu Tan, Siqiao Xue, Xiaoming Shi, James Zhang, and Hongyuan Mei. Bellman meets hawkes: Model-based reinforcement learning via temporal point processes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 9543–9551, 2023.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. 2018.

Rajesh Ranganath, Adler Perotte, Noémie Elhadad, and David Blei. Deep survival analysis. In *Machine Learning for Healthcare Conference*, pp. 101–114. PMLR, 2016.

Pawel Renc, Yugang Jia, Anthony E Samir, Jaroslaw Was, Quanzheng Li, David W Bates, and Arkadiusz Sitek. Zero shot health trajectory prediction using transformer. *NPJ Digital Medicine*, 7 (1):256, 2024.

Alexander Rives, Joshua Meier, Tom Sercu, Siddharth Goyal, Zeming Lin, Jason Liu, Demi Guo, Myle Ott, C Lawrence Zitnick, Jerry Ma, et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences*, 118(15):e2016239118, 2021.

Kjetil Røysland. A martingale approach to continuous-time marginal structural models. *Bernoulli*, 17 (3):895 – 915, 2011.

Kjetil Røysland, Paal Ryalen, Mari Nygaard, and Vanessa Didelez. Graphical criteria for the identification of marginal causal effects in continuous-time survival and event-history analyses. *arXiv preprint arXiv:2202.02311*, 2022.

Helene C Rytgaard, Thomas A Gerds, and Mark J van der Laan. Continuous-time targeted minimum loss-based estimation of intervention-specific mean outcomes. *The Annals of Statistics*, 50(5): 2469–2491, 2022.

Peter Schulam and Suchi Saria. Reliable decision support using counterfactual models. *Advances in neural information processing systems*, 30, 2017.

Nabeel Seedat, Fergus Imrie, Alexis Bellot, Zhaozhi Qian, and Mihaela van der Schaar. Continuous-time modeling of counterfactual outcomes using neural controlled differential equations. *arXiv preprint arXiv:2206.08311*, 2022.

Oleksandr Shchur, Marin Biloš, and Stephan Günnemann. Intensity-free learning of temporal point processes. *arXiv preprint arXiv:1909.12127*, 2019.

Oleksandr Shchur, Ali Caner Türkmen, Tim Januschowski, and Stephan Günnemann. Neural temporal point processes: A review. *arXiv preprint arXiv:2104.03528*, 2021.

Artem Shmatko, Alexander Wolfgang Jung, Kumar Gaurav, Søren Brunak, Laust Hvas Mortensen, Ewan Birney, Tom Fitzgerald, and Moritz Gerstung. Learning the natural history of human disease with generative transformers. *Nature*, 647(8088):248–256, 2025.

Ziyang Song, Qingcheng Lu, He Zhu, David Buckeridge, and Yue Li. Trajgpt: Irregular time-series representation learning for health trajectory analysis. *arXiv preprint arXiv:2410.02133*, 2024.

Ethan Steinberg, Ken Jung, Jason A Fries, Conor K Corbin, Stephen R Pfohl, and Nigam H Shah. Language models are an effective representation learning technique for electronic health record data. *Journal of biomedical informatics*, 113:103637, 2021.

Ethan Steinberg, Jason Fries, Yizhe Xu, and Nigam Shah. Motor: a time-to-event foundation model for structured medical records. *arXiv preprint arXiv:2301.03150*, 2023.

Sindhu Tipirneni and Chandan K Reddy. Self-supervised transformer for sparse and irregularly sampled multivariate clinical time-series. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 16(6):1–17, 2022.

Anastasios Tsiatis. A nonidentifiability aspect of the problem of competing risks. *Proceedings of the National Academy of Sciences*, 72(1):20–22, 1975.

Masatoshi Uehara, Chengchun Shi, and Nathan Kallus. A review of off-policy evaluation in reinforcement learning. *Statistical Science*, 2025.

Utkarsh Upadhyay, Abir De, and Manuel Gomez Rodriguez. Deep reinforcement learning of marked temporal point processes. *Advances in neural information processing systems*, 31, 2018.

Wouter AC van Amsterdam, Pim A de Jong, Joost JC Verhoeff, Tim Leiner, and Rajesh Ranganath. Decision making in cancer: Causal questions require causal answers. *arXiv preprint arXiv:2209.07397*, 2022.

Toon Vanderschueren, Alicia Curth, Wouter Verbeke, and Mihaela van der Schaar. Accounting for informative sampling when learning to forecast treatment outcomes over time. *arXiv preprint arXiv:2306.04255*, 2023.

Yoav Wald, Mark Goldstein, Yonathan Efroni, Wouter AC van Amsterdam, and Rajesh Ranganath. Time after time: Scalable effect estimation for interventions on when and what to do. In *The Thirteenth International Conference on Learning Representations*, 2025.

Zhichao Yang, Avijit Mitra, Weisong Liu, Dan Berlowitz, and Hong Yu. Transformehr: transformer-based encoder-decoder generative model to enhance prediction of disease outcomes using electronic health records. *Nature communications*, 14(1):7857, 2023.

Tianhe Yu, Garrett Thomas, Lantao Yu, Stefano Ermon, James Y Zou, Sergey Levine, Chelsea Finn, and Tengyu Ma. Mopo: Model-based offline policy optimization. *Advances in Neural Information Processing Systems*, 33:14129–14142, 2020.

A Conclusions and Limitation

We developed Competing Event Models, an autoregressive approach for effect estimation under sequential policies in irregular time. Drawing from the competing risks literature, CEMs introduce latent time variables for each event type, enabling straightforward interventions on timing while naturally handling concurrent events via an ordering-based solution. Beyond effect estimation, one can also expand our framework to optimize a timing policy under a reward function. Limitations include the standard no-unobserved-confounding assumption required for identifiability, and the need for rigorous validation on real-world effect estimation questions, both important directions for future work. Replacing the identifiability assumptions, or working on larger problems, may change implementation details. For example, in implementing orderings, the solutions presented in section 4.3 may not be ideal and other approaches can be considered (like using the conversion procedure to produce conversions between representations w.r.t different orderings and teach a model trained on a single ordering how to convert between them). Although accurate causal effect estimation could support clinical decision-making, deployment without sufficient validation risks reinforcing biases present in observational data or producing unreliable recommendations when applied to settings that differ from training.

B Preliminaries and Notation

A summary of notation used throughout the main paper is given in table 2. We consider multivariate temporal point processes (Daley & Vere-Jones, 2003; Andersen et al., 2012) in two settings:

- **Discrete time.** $\mathcal{T}$ is categorical. For simplicity of exposition it is $\mathcal{T} = \{0, 1, \ldots, T\}$ or $\mathcal{T} = \{1, \ldots, T\}$. Time intervals may take values in $\overline{\mathcal{T}}$, and $P(T_{\text{next}}, E_{\text{next}})$ is supported on $\mathcal{E} \times \mathcal{T} \cup (\text{END}, \infty)$. This means $\sum_{t \in \mathcal{T}, e \in \mathcal{E}} P(T_{\text{next}}, E_{\text{next}}) \leq 1$. For CEMs in discrete time, we have that $q(E_{\text{next}} = \text{END} \mid \boldsymbol{T}_{\mathcal{E}} = \boldsymbol{t}_{\mathcal{E}}) = 1$ if and only if $\min_e t_e = \infty$. This follows from the axioms we discuss later, see lemma C.1.
- **No-concurrent-events.** This is the setting discussed in section 4.1 of the main paper. These are processes in continuous time, where time intervals between events take values in $[0, \infty)$ (see e.g. (Andersen et al., 2012) section II.4 for definitions and eq. 2.4.7 for the intensity and section II.7 for the likelihood), and a single event can occur at a certain time. When $N_e(t)$ is the random variable that counts how many events of type e occur until time t, and $N(t) := \sum_e N_e(t)$, then

Table 2: Summary of notation used throughout the paper and appendix.

Symbol	Description
Sets and spaces	
$\mathcal{E}$	Set of event types
$\overline{\mathcal{E}} = \mathcal{E} \cup \{\text{END}\}$	Event types including sequence-end token
$\mathcal{T}$	Time set; continuous $[0, T]$ or finite (discretized/binned)
$\overline{\mathcal{T}} = \mathcal{T} \cup \{\infty\}$	Extended time set
Sequences and histories	
$\mathcal{H} = \{(t_k, e_k)\}_{k=1}^{n}$	Observed trajectory of events
$\mathcal{H}_t = \{(t_k, e_k) \in \mathcal{H} : t_k \leq t\}$	History up to time t.
$\mathcal{H}^J$	Restriction of $\mathcal{H}$ to event types $J \subseteq \mathcal{E}$
$\mathcal{H}^u$	History of unobserved events in $\mathcal{U}$
Distributions and densities	
P	Observational distribution over event sequences
$\tilde{P}$	Interventional distribution
$q_e(T_e \mid \mathcal{H}_t)$	Latent timing distribution for event type e
$\tilde{q}_a(T_a \mid \mathcal{H}_t)$	Interventional timing distribution for event a
$q(E_{\text{next}} \mid \boldsymbol{T}_{\mathcal{E}})$	Tie-breaking / next-event selection distribution
$\lambda_e(t \mid \mathcal{H}_t)$	Intensity (hazard) of event e at time t
$F_e[t \mid \mathcal{H}_t]$	CDF of T_e given history $\mathcal{H}_t$
Random variables and vectors	
$(T_{\text{next}}, E_{\text{next}})$	Time and type of the next event
T_e	Latent clock / timing variable for event type e
$\boldsymbol{T}_{\mathcal{E}} = [T_1, \ldots, T_{\lvert\mathcal{E}\rvert}]$	Vector of latent times for all event types
$\boldsymbol{t}_{\mathcal{E}} = [t_1, \ldots, t_{\lvert\mathcal{E}\rvert}]$	Realization of $\boldsymbol{T}_{\mathcal{E}}$
Special distributions (used in proofs)	
$P_{\boldsymbol{t}_{\mathcal{E}}}$	Point-mass CEM distribution concentrated on $\boldsymbol{t}_{\mathcal{E}}$ (discrete $\mathcal{T}$)
$P_{S_\epsilon(\boldsymbol{t}_{\mathcal{E}})}$	Uniform CEM distribution on ϵ-ball around $\boldsymbol{t}_{\mathcal{E}}$ (continuous $\mathcal{T}$)
$S_\epsilon(\boldsymbol{t}_{\mathcal{E}})$	ϵ-ball $\{\boldsymbol{t}'_{\mathcal{E}} \mid t'_e \in (t_e - \epsilon,\, t_e + \epsilon)\ \forall e\}$
Ordering and unobserved events	
$\prec$	Total ordering over event types in $\mathcal{E}$
$\arg\min_{\prec}$	Arg-min with tie-breaking by $\prec$
$\mathcal{U}$	Set of unobserved event types

this condition is given by $\lim_{h \downarrow 0}[N(t+h) - N(t)] \in \{0, 1\}$ for all t. Here, our convention for ending the sequence is when an event of type END occurs. To match the formalism of the discrete time case, we may push-forward the distribution to assign events of type END time ∞.

We will use distributions of the form $P_{\boldsymbol{t}_{\mathcal{E}}}$ for discrete time or $P_{S_\epsilon(\boldsymbol{t}_{\mathcal{E}})}$ for continuous time, which place all mass on a certain assignment $\boldsymbol{t}_{\mathcal{E}}$ or uniformly over an ϵ-ball around it respectively.

Definition B.1. Let $\boldsymbol{t}_{\mathcal{E}} \in \mathcal{T}^{|\mathcal{E}|}$, where $\boldsymbol{t}_{\mathcal{E}} = [t_1, \ldots, t_{|\mathcal{E}|}]$. When $\mathcal{T}$ is a finite set, we denote by $P_{\boldsymbol{t}_{\mathcal{E}}}$ the distribution under eq. (1) where $q_e(T_e = t_e) = 1 \quad \forall e \in \mathcal{E}$. When $\mathcal{T} = [0, T]$ we denote $P_{S_\epsilon(\boldsymbol{t}_{\mathcal{E}})}$ as the uniform distribution on $S_\epsilon(\boldsymbol{t}_{\mathcal{E}}) = \{\boldsymbol{t}'_{\mathcal{E}} \mid t'_e \in (t_e - \epsilon, t_e + \epsilon)\}$.

For these distributions, $P_{\boldsymbol{t}_{\mathcal{E}}}(E_{\text{next}} = e, T_{\text{next}} = t) = p(e \mid \boldsymbol{T}_{\mathcal{E}} = \boldsymbol{t}_{\mathcal{E}})$, which we write as $p(e \mid \boldsymbol{t}_{\mathcal{E}})$ for short.

C Proofs and Discussions

In this section of the appendix we prove the formal claims in the main paper, and some additional ones to complement a discussion on the details in implementing CEMs and the motivation behind them.

C.1 Proofs of axiomatic characterizations

We begin by proving the axiomatic characterization of $q(E_{\text{next}} \mid \boldsymbol{T}_{\mathcal{E}})$ to choose a minimal time. The axioms are repeated here for convenience.

Axiom 4.1. *For the causal model defined in eq. (1) with $\mathcal{T}$ as a finite set, let the CDFs of $q_e(T_e \mid \mathcal{H}_t)$ be denoted by $F_e[t_e \mid \mathcal{H}_t; \theta] := q_e(T_e < t_e \mid \mathcal{H}_t)$. Let $a \in \mathcal{E}$ and the corresponding interventional distributions of the form $\tilde{P}(\cdot \mid \mathcal{H}_t) = P(\cdot \mid do(\tilde{q}_a(T_a \mid \mathcal{H}_t)))$. For any choice of $\{q_e\}_{e\in\mathcal{E}}$ and an intervention $\tilde{q}_a$, consider the following axioms:*

a. (No backward effect) *If $\tilde{F}_a[s \mid \mathcal{H}_t] = F_a[s \mid \mathcal{H}_t]$ for all $t \le s \le t + \Delta$ where $\Delta > 0$, then $P(E_{\text{next}} = e, T_{\text{next}} = s) = \tilde{P}(E_{\text{next}} = e, T_{\text{next}} = s)$ for all $e \in \mathcal{E}$ and $t \le s < t + \Delta$.*

b. (Expressivity) *Any distribution $P(E_{\text{next}}, T_{\text{next}} | \mathcal{H}_t)$ supported on $\mathcal{E} \times \mathcal{T} \cup (\textit{END}, \infty)$, can be realized by some $\{q_e\}_{e\in\mathcal{E}}$ where $q_e(\infty|\mathcal{H}_t) > 0$ for every $e \in \mathcal{E}$ with $P(E_{\text{next}} = e \mid \mathcal{H}_t) = 0$.*

Lemma C.1. *A model that satisfies expressivity (axiom 4.1(b)) must satisfy $q(E_{\text{next}} = e' \mid \boldsymbol{T}_{\mathcal{E}} = \boldsymbol{t}_{\mathcal{E}}) = 0$ if $t_{e'} = \infty$, and $q(E_{\text{next}} = \textit{END} \mid \boldsymbol{T}_{\mathcal{E}} = \boldsymbol{t}_{\mathcal{E}}) = 1$ if and only if $\min_e t_e = \infty$.*

Proof. To express $P(E_{\text{next}} = \text{END} \mid \mathcal{H}_t) = 1$, the axiom says we must have $q_e(\infty \mid \mathcal{H}_t) > 0$ for all $e \in \mathcal{E}$. Furthermore, because END is the only event that can occur next, we must have $q(\text{END} \mid \boldsymbol{t}_{\mathcal{E}}) = 1$ for any $\boldsymbol{t}_{\mathcal{E}}$ where $q(\boldsymbol{T}_{\mathcal{E}} = \boldsymbol{t}_{\mathcal{E}} \mid \mathcal{H}_t) > 0$. That is because otherwise there is some $e \in \mathcal{E}$ such that $q(e \mid \boldsymbol{t}_{\mathcal{E}}) > 0$, and consequently $P(E_{\text{next}} = e) \ge q(\boldsymbol{t}_{\mathcal{E}} \mid \mathcal{H}_t) q(e \mid \boldsymbol{t}_{\mathcal{E}}) > 0$. Letting $\boldsymbol{t}^*_{\mathcal{E}} = [\infty, \dots, \infty]$, we indeed have $q(\boldsymbol{T}_{\mathcal{E}} = \boldsymbol{t}^*_{\mathcal{E}} \mid \mathcal{H}_t) = \prod_{e\in\mathcal{E}} q_e(\infty \mid \mathcal{H}_t) > 0$ and therefore $q(\text{END} \mid \boldsymbol{t}^*_{\mathcal{E}}) = 1$.

Now, let $\boldsymbol{t}^*_{\mathcal{E}}$ such that $\min_e t^*_e < \infty$, and $F = \{e \in \mathcal{E} \,:\, t^*_e < \infty\}$. We need to show that under the axiom, $q(e \in F \mid \boldsymbol{t}^*_{\mathcal{E}}) = 1$. We let P be any distribution such that $P(E_{\text{next}} \in F \mid \mathcal{H}_t) = 1$ with full support, i.e. $P(e, s \mid \mathcal{H}_t) > 0$ for any $e \in F$ and $s \in \mathcal{T}$.

1. The distribution must be realized by some $\{q_e\}_{e\in\mathcal{E}}$, where $q_e(t^*_e) > 0$ for all $e \in \mathcal{E}$. That is because $q_{e'}(\infty) > 0$ for $e' \notin F$ from the axiom, and also $q_e(t^*_e) > 0$ for $e \in F$ due to the full support of P.
2. Then we have that for $e' \notin F$,

$$0 = P(E_{\text{next}} = e') = \sum_{\boldsymbol{t}_{\mathcal{E}} \in (\mathcal{T}\cup\infty)^{|\mathcal{E}|}} q(e' \mid \boldsymbol{t}_{\mathcal{E}}) \cdot \prod_{e\in\mathcal{E}} q_e(t_e) \ge q(e' \mid \boldsymbol{t}^*_{\mathcal{E}}) \cdot \prod_{e\in\mathcal{E}} q_e(t^*_e).$$

   Because from the previous item we have that $\prod_{e\in\mathcal{E}} q_e(t^*_e) > 0$, we conclude that $q(e' \mid \boldsymbol{t}^*_{\mathcal{E}}) = 0$.

For the continuous case, the same proof holds by replacing $P(\boldsymbol{T}_{\mathcal{E}} = \boldsymbol{t}_{\mathcal{E}}) > 0$ with a requirement for positive measure on neighborhoods of $\boldsymbol{t}_F$ (i.e. the vectors restricted to coordinates $e \in F$ for which $t_e < T$, and are still ∞ at $e \notin F$). The conclusion then holds for almost every $\boldsymbol{t}_{\mathcal{E}}$. □

Lemma 4.1. *For the causal model in eq. (1) to satisfy axiom 4.1 for all settings of $\{q_e\}_{e\in\mathcal{E}}$ and interventions $\tilde{q}_a$ for any $a \in \mathcal{E}$, it must satisfy $q(E_{\text{next}} = e \mid \boldsymbol{T}_{\mathcal{E}} = \boldsymbol{t}_{\mathcal{E}}) = 0 \quad \forall e \notin \arg\min_{e'} t_{e'}$.*

Proof. Consider $P_{\boldsymbol{t}_{\mathcal{E}}}$ for any $\boldsymbol{t}_{\mathcal{E}} \in \overline{\mathcal{T}}^{|\mathcal{E}|}$, assume $t_{\min} := \min_e t_e < \infty$ and let $\mathcal{E}_{\min} = \arg\min_e t_e$. Let $a \notin \mathcal{E}_{\min}$, for an intervention that replaces q_a with $\tilde{q}_a(T_a \mid \mathcal{H}_t) = 1[T_a = \infty]$, we have that under $\tilde{P}$,

$$q(e \mid \tilde{\boldsymbol{t}}_{\mathcal{E}}) = \tilde{P}(T_{\text{next}} = t_e, E_{\text{next}} = e) \overset{(a)}{=} P(T_{\text{next}} = t_e, E_{\text{next}} = e) = q(e \mid \boldsymbol{t}_{\mathcal{E}}) \quad \forall e \in \mathcal{E}_{\min}.$$

The first and third equalities hold because $\tilde{P} = P_{\tilde{\boldsymbol{t}}_\mathcal{E}}$ and $P = P_{\boldsymbol{t}_\mathcal{E}}$ are of the form in definition B.1. The second equality is due to *No effect on past*, axiom (a), because the intervention only changed $F_a[s]$ for $s > t_a > t_e$ and thus $\tilde{F}_a[s \mid \mathcal{H}_t] = F_a[s \mid \mathcal{H}_t] \quad \forall t_a \geq t_e + 1 \geq s \geq t$.

Repeating this for all $a \notin \mathcal{E}_{\min}$ one after the other, we obtain the vector

$$\tilde{\boldsymbol{t}}_\mathcal{E} = \begin{cases} t_e & e \in \mathcal{E}_{\min} \\ \infty & e \notin \mathcal{E}_{\min}. \end{cases}$$

Following the same argument drawn above, we get that $q(e \mid \tilde{\boldsymbol{t}}_\mathcal{E}) = q(e \mid \boldsymbol{t}_\mathcal{E}) \ \forall e \in \mathcal{E}_{\min}$. It also holds from lemma C.1 that $q(e' \mid \tilde{\boldsymbol{t}}_\mathcal{E}) = 0 \ \forall e' \in \mathcal{E} \cup \text{END} \setminus \mathcal{E}_{\min}$. We conclude that $\sum_{e \in \mathcal{E}_{\min}} q(e \mid \boldsymbol{t}_\mathcal{E}) = \sum_{e \in \mathcal{E}_{\min}} q(e \mid \tilde{\boldsymbol{t}}_\mathcal{E}) = 1$, and therefore $q(e' \mid \boldsymbol{t}_\mathcal{E}) = 0 \ \forall e' \notin \mathcal{E}_{\min}$, which proves the claim.

For the continuous case, given that $\min_e t_e < T$ and $\min_{e \notin \mathcal{E}_{\min}} t_e - t_{\min} > 0$ (otherwise the claim is trivial), let $\boldsymbol{t}_\mathcal{E}$ and $0 < \epsilon < |\mathcal{E}|^{-1} \min_{e \notin \mathcal{E}_{\min}} t_e - t_{\min}$. Define the set $S_\epsilon(\boldsymbol{t}_\mathcal{E}) = \{\boldsymbol{t}'_\mathcal{E} | t'_e \in [t_e, t_e + \epsilon]\}$ and consider an intervention $\tilde{q}_a$ that places all mass of the distribution at times later than T, which we defined by convention to be equivalent to ∞ (here again $a \notin \mathcal{E}_{\min}$). Take a distribution $P_{S_\epsilon(\boldsymbol{t}_\mathcal{E})}$ uniform over $S_\epsilon(\boldsymbol{t}_\mathcal{E})$, so it holds that

$$\begin{aligned} q(e \mid \boldsymbol{t}'_\mathcal{E} \in S_\epsilon(\boldsymbol{t}_\mathcal{E})) &= P_{S_\epsilon(\boldsymbol{t}_\mathcal{E})}(e, t \in [t_{\min} - \epsilon, t_{\min} + \epsilon]) \\ &= \tilde{P}(e, t \in [t_{\min} - \epsilon, t_{\min} + \epsilon]) = q(e \mid \boldsymbol{t}'_\mathcal{E} \in \tilde{S}_\epsilon(\boldsymbol{t}_\mathcal{E})) \quad \forall e \in \mathcal{E}_{\min}, \end{aligned}$$

where $\tilde{S}_\epsilon = \{\boldsymbol{t}'_\mathcal{E} \mid t'_e \in [t_e - \epsilon, t_e + \epsilon] \ e \neq a, \ t'_a > T\}$. We also have that $q(a \mid \boldsymbol{t}'_\mathcal{E} \in \tilde{S}_\epsilon(\boldsymbol{t}_\mathcal{E})) = 0$ because $T_a > T$ under $\tilde{S}_\epsilon$. Repeating for all $a \notin \mathcal{E}_{\min}$, the equation above still holds because $|\mathcal{E}| \cdot \epsilon < \min_{e \notin \mathcal{E}_{\min}}$ and therefore the intervention only affects time larger than $t_{\min} + \epsilon$. This means that all mass of $q(e \mid \boldsymbol{t}'_\mathcal{E} \in S_\epsilon(\boldsymbol{t}_\mathcal{E}))$ must be concentrated on the minimizers of $\boldsymbol{t}_\mathcal{E}$. Because this holds for all ϵ, we conclude that the mass must be concentrated on the minimizers for any epsilon ball around $\boldsymbol{t}_\mathcal{E}$, meaning $q(E_{\text{next}} = e \mid \boldsymbol{T}_\mathcal{E} = \boldsymbol{t}_\mathcal{E}) = 0$ for all but a measure zero of time assignments. □

C.2 Expressivity and Correspondence with Intensities in the No-concurrent-events Setting

For multivariate counting processes, their intensities determine the likelihood and distribution of events (Andersen et al., 2012, Chapter II.7). Under our simplifying regularity assumptions, the intensities are defined by $\lambda_e(t \mid \mathcal{H}_t) = \lim_{h \downarrow 0}[N_e(t+h) - N_e(t) = 1 \mid \mathcal{H}_t]$. In the main paper, we claim that CEMs can express any joint distribution and that interventions in this representation are equivalent to interventions on intensities. These statements follow from lemma C.2 below. Given the intensities $\{\lambda_e\}_{e \in \mathcal{E}}$ of a process, it constructs a competing events representation with the same likelihood, which proves the expressivity claim. Then given a CEM $\{q_e\}_{e \in \mathcal{E}}$, we prove an expression of intensity $\lambda_e(t \mid \mathcal{H}_{t-})$ as a functional of q_e. The conclusion from this is that an intervention with $\tilde{q}_a$ results in a distribution $\tilde{P}$ whose intensities λ_e for $e \neq a$ must be identical before and after the intervention.

Lemma C.2. *Let $\mathcal{T} = \mathbb{R}_+$ and consider the distributions $P(T_{\text{next}}, E_{\text{next}} \mid \mathcal{H}_t)$ such that for all $e \in \mathcal{E}$, $P(T_{\text{next}}, E_{\text{next}} = e \mid \mathcal{H}_t)$ is absolutely continuous w.r.t the Lebesgue measure. There is a one-to-one correspondence between competing event representations $\{q_e\}_{e \in \mathcal{E}}$ where the distributions are absolutely continuous w.r.t the Lebesgue measure, and the next event distributions $P(E_{\text{next}}, T_{\text{next}})$.*

Proof. We will show that CEMs are able to express any regular process, and that for a process expressed by a CEM the intensity λ_e is a function only of q_e and not other latent time distributions. The statement follows from the following steps:

- *All processes can be expressed as a CEM:* Let P be a regular process and $\{\lambda_e\}_{e \in \mathcal{E}}$ its intensities, then taking

$$q_e(T_e = s \mid \mathcal{H}_t) = \lambda_e(s \mid \mathcal{H}_t, \mathcal{H}_s = \mathcal{H}_t) \exp\left\{ - \int_t^s \lambda_e(u \mid \mathcal{H}_t, \mathcal{H}_u = \mathcal{H}_t) du \right\},$$

  we get that the probability density for type e jumping at time s, while all the others jumping after that is

$$q(T_{\text{next}} = t, E_{\text{next}} = e \mid \mathcal{H}_t) = q_e(s \mid \mathcal{H}_t) \prod_{e' \neq e} 1 - F_{e'}[s \mid \mathcal{H}_t]$$

$$= \lambda_e(s \mid \mathcal{H}_t, \mathcal{H}_s) \exp\left\{-\int_t^s \lambda_e(u \mid \mathcal{H}_t, \mathcal{H}_u = \mathcal{H}_t)du\right\}$$
$$\prod_{e' \neq e} \exp\left\{-\int_t^s \lambda_{e'}(u \mid \mathcal{H}_t, \mathcal{H}_u = \mathcal{H}_t)du\right\}$$
$$= \lambda_e(s \mid \mathcal{H}_t) \exp\left\{-\int_t^s \lambda_\bullet(u \mid \mathcal{H}_t)du\right\}.$$

Here we used $\lambda_\bullet$ to denote $\sum_{e' \in \mathcal{E}} \lambda_{e'}$. This coincides with the expression for the density of a regular process (Daley & Vere-Jones, 2003, Eq. (7.3.5)).

- *Semantics of interventions.* We have a CEM $\{q_e\}_{e \in \mathcal{E}}$ that induces the process P and an intervention that replaces q_a with $\tilde{q}_a$ to obtain $\tilde{P}$. We will use Q for the joint distribution of a CEM over $T_1, \ldots, T_{|\mathcal{E}|}$. If we denote the last jump time before t by s, then letting $\mathcal{H}_{t-}$ be the history up to time t (not including t), by definition of intensity

$$\begin{aligned}
\lambda_e(t \mid \mathcal{H}_{t-}) &= \lim_{h \to 0} \frac{1}{h} P(N_e(t+h) - N_e(t) = 1 \mid \mathcal{H}_{t-}) \\
&= \lim_{h \to 0} \frac{1}{h} \frac{Q(T_e \in [t, t+h], T_1 \geq t, T_2 \geq t, \ldots, T_{|\mathcal{E}|} \geq t \mid \mathcal{H}_s)}{Q(T_1 \geq t, T_2 \geq t, \ldots, T_{|\mathcal{E}|} \geq t \mid \mathcal{H}_s)} \\
&= -\frac{\partial \log Q(T_1 \geq t_1, T_2 \geq 2, \ldots, T_{|\mathcal{E}|} \geq t_{|\mathcal{E}|} \mid \mathcal{H}_s)}{\partial t_e}\bigg|_{t_1 = \ldots = t_{|\mathcal{E}|} = t} \\
&= -\frac{\partial \log \prod_{e' \in \mathcal{E}} Q_{e'}(T_{e'} \geq t_e \mid \mathcal{H}_s)}{\partial t_e}\bigg|_{t_1 = \ldots = t_{|\mathcal{E}|} = t} \\
&= -\frac{\partial \log Q_e(T_e \geq t_e \mid \mathcal{H}_s)}{\partial t_e}\bigg|_{t_e = t}
\end{aligned} \tag{C.1}$$

Since $Q_e(T_e)$ is only a function of q_e and not q_a for any $e \neq a$, we conclude that an intervention on q_a cannot change any λ_e for $e \neq a$.

□

C.3 Discussion and claims on discrete time and concurrent events

From this point forward, we focus our claims on discretized timelines where $\mathcal{T}$ is categorical. Here, it is also helpful to distinguish between settings with concurrent events and without them.

C.3.1 Discrete time without concurrent events

In this case we consider for simplicity $\mathcal{T} = \{1, 2, \ldots, T\}$, i.e. there is no time 0 which marks concurrency. Considering CEMs in this setting, the main difference from the absolutely continuous case is that there is a strictly positive probability that the minimizer of $\boldsymbol{T}_\mathcal{E}$ is non-unique. This means that $q(E_{\text{next}} \mid \boldsymbol{T}_\mathcal{E})$ must determine how to break ties between minimizers. We make two claims on this:

1. Consider interventions $P(\cdot \mid do(\tilde{q}_a(T_a \mid \mathcal{H}_t)))$ on $a \in \mathcal{E}$. There is no single tie-breaking strategy under which interventions on *each* $a \in \mathcal{E}$ have an intuitive meaning.
2. For a specific $a \in \mathcal{E}$, under a tie-breaking strategy where a is never chosen when it is tied for the minimum time, the interventions have intuitive semantics. Specifically, they are synonymous with interventions on the intensity λ_a and can express any such intervention.

The requirement for tie-breaking strategies that change according to the intervened event motivates us to consider orderings $\prec$. There are two ways to reason about this claim. The first does not appeal to intensities, the second takes interventions on intensities as the correct objects to mimic with competing event representations, and shows the equivalence between interventions in both representations.

Axiomatic claim. Consider adding the following to axiom 4.1.

Axiom C.1. *[No instanteneous effect] If $\tilde{F}_a[s \mid \mathcal{H}_t] = F_a[s \mid \mathcal{H}_t]$ for all $t \leq s \leq t_0$, then $P(E_{\text{next}} = e, T_{\text{next}} = t_0) = \tilde{P}(E_{\text{next}} = e, T_{\text{next}} = t_0)$ for all $e \neq a$.*

The axiom extends axiom 4.1(a) to postulate that the effect of a T_a intervention on the probabilities of other events $e \neq a$ cannot be instantaneous. That is, if we only change the distribution of a's clock from time t_0 forward, it does not affect the probability of other events occurring at t_0 itself. Imagine that a corresponds to prescribing statins, and we ask what happens if no statins are given after $t_0 = 4$ weeks (and up until then we prescribe them as in the observed distribution). So the probability of other events, e.g. MI, to occur *exactly* at 4 weeks should not change, because we "prevent" people from taking statins at time t_0 and for that to make an effect on MI, we have to wait at least one time unit until $t_0 + 1$. That is, changing a's clock at time t_0 can only affect other events at time $t_0 + 1$. A similar assumption is also baked into most formulations of counting processes, when stating that intensities are predictable from the filtration $\mathcal{H}_{t-}$, which in our notation means the history up until time t, t not included.

Let us show that for axiom C.1 to hold, the action a must not be chosen when t_a it is the non-unique minimizer of $\boldsymbol{t}_\mathcal{E}$ (or in other words, it should only be chosen when t_a is the unique minimizer). Intuitively, the result holds because if an event a can win a tie, then increasing/decreasing $q_a(t_0)$ will decrease/increase the probability of other events e that may lose to a under a tie at t_0. If a cannot win a tie, then such a change cannot happen.

Lemma C.3. *Under the causal model in eq. (1) where $\mathcal{T}$ is categorical, it holds that:*

1. *A distribution that satisfies axiom 4.1 and axiom C.1 for interventions on a specific $a \in \mathcal{E}$, the distribution $q(E_{\text{next}} \mid \boldsymbol{T}_\mathcal{E})$ must satisfy $q(a \mid \boldsymbol{t}_\mathcal{E}) = 1[a = \arg\min_{e'} t_{e'}]$.*
2. *There is no single distribution $q(E_{\text{next}} \mid \boldsymbol{T}_\mathcal{E})$ that satisfies axiom 4.1 and axiom C.1 for interventions $\tilde{q}_e$ on all $e \in \mathcal{E}$, i.e. not on a specific $a \in \mathcal{E}$ (and choices of $\{q_e\}_{e \in \mathcal{E}}$).*

Proof. The second item, that says no single distribution $q(E_{\text{next}} \mid \boldsymbol{T}_\mathcal{E})$ satisfies the axioms for interventions on all $e \in \mathcal{E}$, is an immediate consequence of the first item. This is because if $\boldsymbol{t}_\mathcal{E}$ has non-unique minimizers, say e_1, e_2, then to have the interventions on both e_1, e_2 satisfy the axioms, the first item says that both must lose the competition (i.e. have probability 0 of being chosen in $q(E_{\text{next}} = e \mid \boldsymbol{T}_\mathcal{E} = \boldsymbol{t}_\mathcal{E}) = 0$ for $e \in \{e_1, e_2\}$). But then no event can win and $q(E_{\text{next}} \mid \boldsymbol{T}_\mathcal{E} = \boldsymbol{t}_\mathcal{E})$ is not a distribution. Let's prove the first item.

Lemma 4.1 tells us that to satisfy axiom 4.1, the distribution must give $\sum_{e \in \arg\min t_e} q(e \mid \boldsymbol{t}_\mathcal{E}) = 1$ for any $\boldsymbol{t}_\mathcal{E}$ that has an entry smaller than ∞. Specifically, if $a \notin \arg\min_e t_e$ then $q(a \mid \boldsymbol{t}_\mathcal{E}) = 0$, and if $a = \arg\min_e t_e$ (i.e. it is the unique minimizer) then $q(a \mid \boldsymbol{t}_\mathcal{E}) = 1$. It is left to show for the case where a is a non-unique minimizer let $\tilde{\boldsymbol{t}}_\mathcal{E}$ such that $|\arg\min_e \tilde{t}_e| > 1 \Rightarrow q(a \mid \tilde{\boldsymbol{t}}_\mathcal{E}) = 0$.

To this end, let $\tilde{\boldsymbol{t}}_\mathcal{E}$ be a vector of times that satisfy this. Take any $\tilde{t}_a < t_a$ (note that such t_a must exist because $\tilde{t}_a$ is minimal and $\tilde{\boldsymbol{t}}_\mathcal{E}$ has entries smaller than ∞) and construct $\boldsymbol{t}_\mathcal{E}$ such that

$$t_e = \begin{cases} \tilde{t}_e & e \neq a \\ t_a & e = a. \end{cases}$$

For $\boldsymbol{t}_\mathcal{E}$ denote $t_{\min} = \min_e t_e$ and note that $a \notin \arg\min_e t_e$. Considering the distribution $P_{\boldsymbol{t}_\mathcal{E}}$ and intervention $\tilde{q}_a(T_a) = 1[T_a = \min_e t_e]$. That is, the intervention shifts $P_{\boldsymbol{t}_\mathcal{E}}$ to $P_{\tilde{\boldsymbol{t}}_\mathcal{E}}$. Then from *No instantaneous effect*, we get

$$\begin{aligned} q(e \mid \tilde{\boldsymbol{t}}_\mathcal{E}) = &\tilde{P}(E_{\text{next}} = e, T_{\text{next}} = t_{\min}) \overset{(axiom\ C.1)}{=} \\ &P_{\boldsymbol{t}_\mathcal{E}}(E_{\text{next}} = e, T_{\text{next}} = t_{\min}) = q(e \mid \boldsymbol{t}_\mathcal{E}) \quad \forall e \in \arg\min_e t_e, \end{aligned}$$

where the use of the axiom is valid because $F_a[s \mid \mathcal{H}_t] = \tilde{F}_a[s \mid \mathcal{H}_t]$ for $t \leq s \leq t_{\min}$. As a consequence we get that $\sum_{e \in \arg\min_e \tilde{t}_e} q(e \mid \tilde{\boldsymbol{t}}_\mathcal{E}) = 1$, and because $a \notin \arg\min_e t_e$, we conclude that $q(a \mid \tilde{\boldsymbol{t}}_\mathcal{E}) = 0$ as we claimed.

□

Equivalence to interventions on intensities. The second method to see that no single tie-breaking technique results is a causal model that satisfies all the axioms for all interventions, is through the relation between interventions on intensities and interventions on latent times.

Intensities in discrete time are defined by $\lambda_e(t_{\text{next}}) = P(t_{\text{next}}, e)/P(T_{\text{next}} \geq t_{\text{next}})$, and we will call the representation $\{\lambda_e\}_{e\in\mathcal{E}}$ the intensity based representation of $P(T_{\text{next}}, E_{\text{next}})$. In continuous time, it is common to define interventions on time as replacing $\lambda_e \to \tilde{\lambda}_e$, and we can discuss such interventions in discrete time as well.

Note: We can view this as a process where $\lambda(t+1 \mid \mathcal{H}_t)$ is a distribution over what happens at time $t+1$, conditioned on history up until time t. The components of this distribution are $\lambda_e(t+1 \mid \mathcal{H}_t)$ and there is an added event $\emptyset$ that means nothing happened at time $t+1$. The distributions $\lambda_e(s \mid \mathcal{H}_t)$ for $s > t+1$ are again the distributions on what happens at time s, but conditioned that the event $\emptyset$ occurred at all previous times $t < u < s$. Then the intervention on λ_e for $e \neq \emptyset$ shifts mass between e and $\emptyset$.

Seeing that interventions in competing events representations $q_a \to \tilde{q}_a$ and in intensity based representations $\lambda_a \to \tilde{\lambda}_a$ are both supposed to describe interventions on time, then we should expect some equivalence between them. Given an intensity based representation λ and an intervention $\tilde{\lambda}$ where $\lambda_e = \tilde{\lambda}_e \ \forall e \neq a$, we expect there will be two competing event representations $q, \tilde{q}$ where: *(i)* q, λ represent the same joint distribution $P(T_{\text{next}}, E_{\text{next}})$ and likewise $\tilde{q}, \tilde{\lambda}$ represent the same $\tilde{P}(T_{\text{next}}, E_{\text{next}})$; *(ii)* $q_e = \tilde{q}_e \ \forall e \neq a$. Similarly to our claims with the no-instantaneous-effect axiom, we show that $q(E_{\text{next}} \mid \boldsymbol{T}_{\mathcal{E}})$ needs to be adjusted, according to the event $e \in \mathcal{E}$ that we intervene on. We also show that once we set this distribution so that the intervened event $a \in \mathcal{E}$ never wins a tie, then any intervention on q_a corresponds to a intervention on λ_a and any intervention on λ_a has an equivalent intervention in a competing-event-representation (specifically, we will take an ordering, see section 4.2, and prove this for orderings where a is last).

First, the following lemma gives a counterexample showing that the tie-breaking strategy $q(E_{\text{next}} \mid \boldsymbol{T}_{\mathcal{E}})$ has to be adjusted to the intervened event a.

Lemma C.4. *There exists a joint distribution $P(T_{\text{next}}, E_{\text{next}})$ with derived intensities λ_e, such that for any tie-breaking strategy p_{tie} and CEM that induce the distribution P, interventions on q_a do not coincide with interventions on λ_a.*

Proof. Consider $|\mathcal{T}| = 1$ and $\mathcal{E} = \{1, 2\}$ and let $\{q_e\}_{e\in\mathcal{E}}$ be a CEM where

	$t=1$	$t=\infty$
q_1	α	$1-\alpha$
q_2	β	$1-\beta$

Note that here a tie breaking strategy can be fully represented by a scalar $\gamma := q(E_{\text{next}} = 1 \mid \boldsymbol{T}_{\mathcal{E}} = [1, 1])$ and then the induced joint distribution $P(E_{\text{next}}, T_{\text{next}})$ is

$$P(E_{\text{next}}, T_{\text{next}}) = \begin{array}{|c|c|}\hline E_{\text{next}} \backslash T_{\text{next}} & t=1 \\ \hline 1 & \alpha(1-\beta) + \gamma\alpha\beta \\ \hline 2 & \beta(1-\alpha) + (1-\gamma)\alpha\beta \\ \hline \end{array} \tag{C.2}$$

It is easily verified that the probability of no event occurring is $(1-\alpha)(1-\beta)$. In this simple case of a single time slot, the intensity exactly equals the joint probability, i.e. $\lambda_e(1) = P(e, 1)$. Assume w.l.o.g that $\alpha, \beta > 0$ and consider the intervention on q that replaces α by $\alpha/2$, and the one that replaces β with $\beta/2$. The joint distributions they induce are:

$$\tilde{P}^1(E_{\text{next}}, T_{\text{next}}) = \begin{array}{|c|c|}\hline E_{\text{next}} \backslash T_{\text{next}} & t=1 \\ \hline 1 & \alpha/2(1-\beta) + \gamma\alpha\beta/2 \\ \hline 2 & \beta(1-\alpha/2) + (1-\gamma)\alpha\beta/2 \\ \hline \end{array} \tag{C.3}$$

$$\tilde{P}^2(E_{\text{next}}, T_{\text{next}}) = \begin{array}{|c|c|}\hline E_{\text{next}} \backslash T_{\text{next}} & t=1 \\ \hline 1 & \alpha(1-\beta/2) + \gamma\alpha\beta/2 \\ \hline 2 & \beta/2(1-\alpha) + (1-\gamma)\alpha\beta/2 \\ \hline \end{array} \tag{C.4}$$

Since the first intervention only changes q_1 and the second changes q_2, we would like the second row of C.3 to equal the second row in C.2. Likewise the first row of C.4 should equal the first row of C.2. These desired properties are expressed by the following two equations being satisfied,

$$\beta(1-\alpha/2) + (1-\gamma)\alpha\beta/2 = \beta(1-\alpha) + (1-\gamma)\alpha\beta$$

$$\alpha(1-\beta/2)+\gamma\alpha\beta/2=\alpha(1-\beta)+\gamma\alpha\beta$$

Moving terms that include γ to the left hand-side in both equations gives us the following system,

$$\begin{aligned}\gamma\alpha\beta/2&=0\\-\gamma\alpha\beta/2&=-\alpha\beta/2\end{aligned}$$

Since α,β are non-zero positive numbers, we see that the first equation requires $\gamma=0$ (i.e. event 2 always wins) and the second requires $\gamma=1$, hence there is no CEM where interventions on q_e coincide with those on λ_e. □

The requirement for $\gamma=0$ or $\gamma=1$ at the end of the above proof hints at the requirement for the intervened event to always lose a tie. In the next result, under the notation of orderings that we introduce in section 4.2 of the main paper, it is shown that when tie-breaking satisfies this requirement, the expected correspondence between intensities and competing events holds indeed. The result is built up of 3 short claims, the last of which being theorem 5.1 from the main paper, which shows the equivalence of interventions in both representations as discussed above.

Lemma C.5. *Let (q,λ) and $(\tilde{q},\tilde{\lambda})$ be representations of a distribution over the next time and event, before and after an intervention on q_a respectively. Further assume that ties in $q,\tilde{q}$ are broken using an ordering $\succ$, then $\lambda_e=\tilde{\lambda}_e\ \forall e\prec a$.*

Proof. Let $e\prec a$ and $t\in\mathcal{T}$, transforming q to a joint probability $P(E_{\text{next}},T_{\text{next}})$ we get that

$$P(e,t)=q_e(t)\prod_{e'\prec e}\left[\sum_{k\geq t+1}q_{e'}(k)\right]\prod_{e\prec e'}\left[\sum_{k\geq t}q_{\tilde{e}}(k)\right],$$

while the survivor function $S(t)$ is simply the probability that all competing events jumped at time t or later, given by

$$S(t)=\prod_{e'\in\mathcal{E}}\left[\sum_{k\geq t}q_{e'}(k)\right]$$

Denoting by $\tilde{P}$ and $\tilde{S}$ the corresponding functions under $\tilde{q}$, i.e. post-intervention on q_a, we have that:

$$\begin{aligned}\tilde{\lambda}_e(t)=\frac{\tilde{P}(e,t)}{\tilde{S}(t)}&=\frac{\tilde{q}_e(t)\prod_{e'\prec e}\left[\sum_{k\geq t+1}\tilde{q}_{e'}(k)\right]\prod_{e\prec e'}\left[\sum_{k\geq t}\tilde{q}_{e'}(k)\right]}{\prod_{e'}\left[\sum_{k\geq t}\tilde{q}_{e'}(k)\right]}\\&=\frac{\tilde{q}_e(t)}{\sum_{k\geq t}\tilde{q}_e(t)}\frac{\prod_{e'\prec e}\left[\sum_{k\geq t+1}\tilde{q}_{e'}(k)\right]}{\prod_{e'\prec e}\left[\sum_{k\geq t}\tilde{q}_{e'}(k)\right]}\cdot\frac{\prod_{e\prec e'}\left[\sum_{k\geq t}\tilde{q}_{e'}(k)\right]}{\prod_{e\prec e'}\left[\sum_{k\geq t}\tilde{q}_{e'}(k)\right]}\\&=\frac{q_e(t)}{\sum_{k\geq t}q_e(t)}\frac{\prod_{e'\prec e}\left[\sum_{k\geq t+1}q_{e'}(k)\right]}{\prod_{e'\prec e}\left[\sum_{k\geq t}q_{e'}(k)\right]}\cdot\frac{\prod_{e\prec e'}\left[\sum_{k\geq t}q_{e'}(k)\right]}{\prod_{e\prec e'}\left[\sum_{k\geq t}q_{e'}(k)\right]}\\&=\frac{q_e(t)\prod_{e'\prec e}\left[\sum_{k\geq t+1}q_{e'}(k)\right]\prod_{e\prec e'}\left[\sum_{k\geq t}q_{e'}(k)\right]}{\prod_{e'}\left[\sum_{k\geq t}q_{e'}(k)\right]}=\lambda_e(t)\end{aligned}$$

The third line, where we replace $\tilde{q}$ by q holds because the right item in the multiplication equals 1 and we write it with densities q instead of $\tilde{q}$, while the left item in the multiplication only contains $\tilde{q}_{e'}$ where $e'\prec e$, i.e. these types e' precede the intervened process since $e'\prec e\prec a$ and thus $e'\neq a$ and $\tilde{q}_{e'}=q_{e'}$. The above equality proves our claim, and we also see that for $e\succ a$ the expression would include the ratio $\sum_{k\geq t+1}\tilde{q}_a(k)/\sum_{k\geq t}\tilde{q}_a(k)$, which changes under an intervention on q_a, hence these intensities are not invariant to the intervention. □

Lemma C.6. *Let (q, λ) be a representation of a distribution over the next time and event, assume ties in q are broken using an ordering $\succ$ and that e is last in the ordering. Then we have*

$$\lambda_a(t) = \left(1 - \sum_{e \neq a} \lambda_a(t)\right) \frac{q_a(t)}{\sum_{t' \geq t} q_a(t')}. \tag{C.5}$$

Proof. Since a is last in the ordering, we have the following equality:

$$\begin{aligned}\lambda_a(t) = \frac{P(a,t)}{S(t)} &= \frac{q_a(t) \prod_{e \neq a} \left[\sum_{k \geq t+1} q_e(k)\right]}{\prod_{e \in \mathcal{E}} \left[\sum_{k \geq t} q_e(k)\right]} \\ &= \frac{q_a(t)}{\sum_{k \geq t} q_a(k)} \cdot \frac{\prod_{e \neq a} \left[\sum_{k \geq t+1} q_e(k)\right]}{\prod_{e \neq a} \left[\sum_{k \geq t} q_e(k)\right]}.\end{aligned} \tag{C.6}$$

Now we will show that the right item in the resulting multiplication is equal to $1 - \sum_{e \neq a} \lambda_e(t)$. Begin with

$$\begin{aligned}1 - \sum_{e \neq a} \lambda_e(t) &= \frac{S(t) - \sum_{e \neq a} P(e,t)}{S(t)} \\ &= \frac{\prod_{e \in \mathcal{E}} \sum_{k \geq t} q_e(k) - \sum_{e \neq a} q_e(t) \prod_{e' \prec e} \left[\sum_{k \geq t+1} q_{e'}(k)\right] \prod_{e \prec e'} \left[\sum_{k \geq t} q_{e'}(k)\right]}{\prod_{e \in \mathcal{E}} \sum_{k \geq t} q_e(k)} \\ &= \frac{\prod_{e \neq a} \sum_{k \geq t} q_e(k) - \sum_{e \neq a} q_e(t) \prod_{e' \prec e} \left[\sum_{k \geq t+1} q_{e'}(k)\right] \prod_{\substack{e \prec e', \\ e' \neq a}} \left[\sum_{k \geq t} q_{e'}(k)\right]}{\prod_{e \neq a} \sum_{k \geq t} q_e(k)}.\end{aligned} \tag{C.7}$$

In the first line we simply switch in the definition of $\lambda_e(t)$ in terms of joint probabilities and the survival function. Then in second line we replace it with the latent time distributions. In the third line we pull out the items in the product where $e = a$, note that the equality is correct because e is last in the ordering and therefore the same expression, $\prod_{e \prec e'} [\sum_{k \geq t} q_{e'}(k)]$, appears in all the products when e' takes value a. It is left to show that the numerator above equals $\prod_{e \neq a} \left[\sum_{k \geq t+1} q_e(k)\right]$, then we will plug the entire expression into the one we derived for $\lambda_a(t)$ in eq. (C.5) and conclude the proof. To this end let us assign notations to the event types that correspond to the ordering $\succ$, i.e. $e_1 \prec e_2 \ldots \prec e_{|\mathcal{E}|-1}$.[8] For any ordered set of events we have the following identity:

$$\begin{aligned}&\prod_{j=1}^{|\mathcal{E}|-1} \sum_{k \geq t} q_{e_j}(k) - \prod_{j=1}^{|\mathcal{E}|-1} \sum_{k \geq t+1} q_{e_j}(k) \\ &= \left(q_{e_1}(t) + \sum_{k \geq t+1} q_{e_1}(k)\right) \prod_{j=2}^{|\mathcal{E}|-1} \sum_{k \geq t} q_{e_j}(k) - \prod_{j=1}^{|\mathcal{E}|-1} \sum_{k \geq t+1} q_{e_j}(k) \\ &= q_{e_1}(t) \cdot \prod_{j=2}^{|\mathcal{E}|-1} \sum_{k \geq t} q_{e_j}(k) - \left(\sum_{k \geq t+1} q_{e_1}(k)\right) \left(\prod_{j=2}^{|\mathcal{E}|-1} \sum_{k \geq t} q_{e_j}(k) - \prod_{j=2}^{|\mathcal{E}|-1} \sum_{k \geq t+1} q_{e_j}(k)\right).\end{aligned}$$

The rightmost item in the expression is identical to the expression we began with, but with the starting index on the products raised to 2. Continuing by induction we get that

$$\prod_{j=1}^{|\mathcal{E}|-1} \sum_{k \geq t} q_{e_j}(k) - \prod_{j=1}^{|\mathcal{E}|-1} \sum_{k \geq t+1} q_{e_j}(k) = \sum_{j=1}^{|\mathcal{E}|-1} q_{e_j}(t) \prod_{l=1}^{j-1} \left[\sum_{k \geq t+1} q_{e_l}(k)\right] \prod_{l=j+1}^{|\mathcal{E}|-1} \left[\sum_{k \geq t} q_{e_l}(k)\right]$$

[8] Here we will have $a = e_{|\mathcal{E}|}$ since we assumed e is last in the ordering, yet we omit it since we do not use a in the rest of the derivation.

$$= \sum_{e \neq a} q_e(t) \prod_{e' \prec e} \left[\sum_{k \geq t+1} q_{e'}(k) \right] \prod_{\substack{e \prec e', \\ e' \neq a}} \left[\sum_{k \geq t} q_{e'}(k) \right]$$

Plugging this into eq. (C.7), we get

$$1 - \sum_{e \neq a} \lambda_e(t) = \frac{\prod_{e \neq a} \left[\sum_{k \geq t+1} q_e(k) \right]}{\prod_{e \neq a} \left[\sum_{k \geq t} q_e(k) \right]},$$

and plugging this into eq. (C.6) completes the proof. □

Theorem 5.1. *Let λ and $\tilde{\lambda}$ be intensities that represent distributions over the next time and event, before and after an intervention on λ_a. There exists competing event models $q, \tilde{q}$ and ordering $\prec$ under which the models induce the same distributions as $\lambda, \tilde{\lambda}$ respectively, and $q_e = \tilde{q}_e \; \forall e \neq a$.*

Proof. Let $\succ$ be any ordering under which a is last. Note that because $\sum_e \lambda_e(t) \leq 1$ (due to the definition of intensities) we must have $\tilde{\lambda}_a(t) \leq 1 - \sum_{e \neq a} \lambda_e(t)$. In other words $\tilde{\lambda}_a(t) = c \cdot \left(1 - \sum_{e \neq a} \lambda_e(t)\right)$ for some $c \in [0, 1]$. To conclude the proof, let us construct the transformation from λ to $\tilde{\lambda}$.
We initialize $\tilde{q}_a = q_a$. For $t = T \dots 1$ do:

- Take the value $c \in [0, 1]$ for which $\tilde{\lambda}_a(t) = c \cdot \left(1 - \sum_{e \neq a} \lambda_e(t)\right)$.
- Update $\tilde{q}_a(t') \leftarrow (1 - c) \cdot \tilde{q}_a(t')$ for all $t' > t$.
- Update $\tilde{q}_a(t) \leftarrow c \cdot \sum_{t' \geq t} \tilde{q}_a(t')$.

The construction only sets $\tilde{q}_a$ and not $\tilde{q}_e$ for any $e \neq a$. Therefore from lemma C.5, the distribution defined by $\tilde{q}$ has intensities $\tilde{\lambda}_e = \lambda_e$ for all $e \neq a$. It is also easy to see that at each step we set $\tilde{q}$ such that $\tilde{q}_a(t) / \sum_{t' \geq t} \tilde{q}_a(t') = c$, for the appropriate c defined at that step. This relation is kept under the following iterations as we iterate backward in $t = T \dots 1$, since at the second step $\tilde{q}_a(t')$ are always multiplied by the same scalar for all $t' > t$. From lemma C.6 this shows that the intensity of event a in the distribution defined by $\tilde{q}$ is indeed the desired $\tilde{\lambda}_a$ and concludes our proof. □

C.3.2 Semantics of interventions under concurrent events

This section discusses the differences that arise between interventions on intensities and those on competing events, when applied to concurrent events. Under a naïve solution where q_J or λ_J for each $J \subseteq \mathcal{E}$ are added to account for concurrent events, the correspondence of the last section applies as it did there. For efficient implementation, we are interested in interventions when $\mathcal{T} = \{0, 1, \dots, T\}$, so $t_{\text{next}} = 0$ denotes concurrent events.

When there are concurrent events, intensities also require some tie-breaking-like mechanism, as the competing event representations do. That is because, as we mentioned in section 4.2, autoregressive generation of the event $J = \{e'_1, \dots, e'_L\}$ can be generated as $(t, e'_1), (0, e'_2), \dots, (0, e'_L)$ or equivalently as $(t, e'_{\sigma(1)}), (0, e'_{\sigma(2)}), \dots, (0, e'_{\sigma(L)})$ for some permutation on $[L]$. So again we can consider an ordering, where one event is always generated first between the L items, or some more general stochastic mechanism. Let us start by considering an ordering based strategy like we do for competing events, where one event is always generated before the other.

Intuitively, an intervention on timing should be interpreted as changing *the probability that event e happens at time s, conditioned on nothing happening before time s*. When we fully formalize this statement, there is an ambiguity with concurrent events: should we condition on any information about what happens at time s itself?
Without concurrent events, there is no ambiguity, and $\lambda_e(s)$ is exactly this quantity we wish to intervene on. $q_e(s)$ on the other hand, can be described as *the probability that event e happens at time s, conditioned on no event $e' \neq e$ happening before time s and no events $e' \prec e$ happening at time s itself.* This differs from $\lambda_e(s)$, in two ways:

1. $q_e(s)$ conditions on no event $e' \neq e$ happening before time s, whereas $\lambda_e(s)$ also conditions on e itself not happening before time s. This changes the normalization of probabilities, so $q_e(\bullet)$ is a distribution over time, while $\lambda_\bullet(s)$ is a distribution over event types. This makes little difference for the semantics of interventions, but has some consequences for implementation when we model these quantities.
2. What matters when we interpret interventions on $q_e(s)$, is that it also conditions on $e' \prec e$ not happening at time s itself. Notice that when e is last under $\prec$ then $q_e(s)$ conditions on no events $e' \neq e$ happening in time s itself. This meaning aligns with $\lambda_e(s)$, which explains why theorem 5.1 holds and the construction uses the last event under $\prec$. We say the quantities align because $\lambda_\bullet(s)$ is a distribution over events, which means the occurrence of one event is mutually exclusive from the other. $q_e(s)$ for e last under $\prec$, enforces this mutual exclusivity through conditioning, whereas for $e' \neq e$ this alignment does not happen.

   Conclusion: When there are no concurrent events, conditioning on $e' \prec e$ not happening at time s can result in unintuitive interventions when e is not last under $\prec$. Vice-versa, under concurrent events interventions on q_e are more intuitive than those on λ_e, which do not condition on any information about what happens at time s. We illustrate this intuition with an example below. We will return to the conclusion after the examples.

It is instructive to consider what happens with two events that are independent, for example, e_1, e_2 are prescriptions for statins and prescriptions for antibiotics given for pneumonia (such as Amoxicillin) accordingly. Let us reiterate the example in lemma C.4, where we had $\mathcal{T} = \{1\}$ and write down the probability tables we had there. For a simple illustration, we'll set arbitrary probabilities for e_1, e_2, say 0.1 and 0.3 accordingly.

First, the distribution in terms of intensities is equal to the joint probabilities:

$$\lambda_{E_{\text{next}}}(T_{\text{next}}) = P(E_{\text{next}}, T_{\text{next}}) =$$

E_{next} \ T_{next}	$t = 1$
1	0.1
2	0.3

If we intervene and set $\tilde{\lambda}_{e_1}(1) = 0.2$, then as expected the probability of e_2 does not change.

$$\tilde{\lambda}_{E_{\text{next}}}(T_{\text{next}}) =$$

E_{next} \ T_{next}	$t = 1$
1	0.2
2	0.3

The extra 0.1 probability mass for e_1 is reduced from the probability that nothing happens at the timestep $t = 1$. Here, the "independence" of 2 events is their mutual exclusivity.

Writing the pre-intervention distribution in its competing events representation, under the ordering $e_1 \prec e_2$ we get:

	$t = 1$	$t = \infty$
q_1	0.1	0.9
q_2	$0.3/(1 - 0.1)$	$0.7/(1 - 0.1)$

,

	$t = 1$	$t = \infty$
$\tilde{q}_1$	0.2	0.8
q_2	$0.3/(1 - 0.1)$	$0.7/(1 - 0.1)$

Here, raising the probability of e_1 to 0.2 also reduces the probability that e_2 wins. The added probability mass is not reduced only from the probability that nothing happens, but also from e_2. The corresponding joint probabilities are:

$$P(E_{\text{next}}, T_{\text{next}}) =$$

E_{next} \ T_{next}	$t = 1$
1	0.1
2	0.3

$$, \tilde{P}(E_{\text{next}}, T_{\text{next}}) =$$

E_{next} \ T_{next}	$t = 1$
1	0.2
2	$4/15$

.

This is counterintuitive because if we think the two treatments are mechanistically independent, then raising the probability of one should not repress the other. In an autoregressive model, we might retrieve this lost probability mass by conditioning on the first event (which after intervention, will have more e_1 events generated) and then perhaps generating e_2 events in the next step. But the times of these e_2 events will no longer be time 1 since we do not allow concurrent events.

When concurrent events are allowed, the type of independence that intensities give is not necessarily the one we seek. Here, it is interesting to observe what happens under statistical independence between e_1 and e_2. Say that each treatment happens at time $t = 1$ as before with probabilities 0.1 and 0.3, and they are independent, so they co-occur with probability $0.1 \cdot 0.3 = 0.03$. Under an intervention that raises the probability of e_1 to 0.2, we expect the probability of $\{e_1, e_2\}$ occurring at $t = 1$ to be $0.2 \cdot 0.3 = 0.06$ To generate the co-occuring event we may need two autoregressive steps. In the competing events representation with ordering $e_1 \prec e_2$, the two distributions we need to specify are as follows:

	$t=0$	$t=1$	$t=\infty$
q_1	0	0.1	0.9
q_2	0	0.3	0.7

, $q_2(0 \mid (1, e_1)) = 0.3$

The first distribution just assigns the probability of an event happening at time $t = 1$, then if the first event was e_1 then according to the ordering, e_2 might be generated with it, and this is why we have to specify $q_2(T_2 = 0 \mid (1, e_1))$. These probabilities full specify the distribution of the first event. It is easy to calculate that the above probabilities indeed express the distribution of interest. Under the intervention on e_1, the distributions change intuitively:

	$t=0$	$t=1$	$t=\infty$
$\tilde{q}_1$	0	**0.2**	**0.8**
q_2	0	0.3	0.7

, $q_2(0 \mid (1, e_1)) = 0.3$.

The probability of co-occuring events is $0.2 \cdot 0.3$, while the probabilities for each event (either occurring alone or together with the other) are 0.2 and 0.3 respectively.

Under the intensities representation, the probabilities do not change intuitively. As we mentioned earlier, with concurrent events an ordering for the intensity based representation is also needed. We assume again e_1 is always generated before e_2, it is easy to show that the intensities-based representation is

$\lambda_{E_{\text{next}}}(T_{\text{next}}) =$

E_{next} \ T_{next}	$t=0$	$t=1$
1	0	0.1
2	0	0.27

, $\lambda_{e_2}(T_2 = 0 \mid (1, e_1)) = 0.3$.

Now, under intervention the intensities are

$\tilde{\lambda}_{E_{\text{next}}}(T_{\text{next}}) =$

E_{next} \ T_{next}	$t=0$	$t=1$
1	0	**0.2**
2	0	0.27

, $\lambda_{e_2}(T_2 = 0 \mid (1, e_1)) = 0.3$.

The probability that e_2 occurs (either jointly with e_1 or not) is $0.27 + 0.2 \cdot 0.3 = 0.33$, so the intervention raised it from its original value of 0.3. Furthermore, the probability of a concurrent event $(\{e_1, e_2\}, 1)$ is $0.2 \cdot 0.3 = 0.06$. But the product of probabilities of both events is $0.2 \cdot 0.33 \neq 0.06$, so the intervention on e_2 both changed the probability of e_1 and created a correlation between the occurrence of the two events. Both these consequences are not intuitive when we think about interventions on independent events. The underlying calculation here is that the intensity is obtained when we divide the joint probability by the survival function $S(t)$, so we in fact condition on no events happening beforehand. For the competing event type intervention (which is arguably more intuitive in this case), we should've set a different survival function for λ_2, i.e. the probability that no event happened beforehand, and e_1 did not jump at $t = 1$. But if we do that, then $\lambda_\bullet(t)$ will no longer be a distribution over events (with leftover probability that nothing happens). For example, it might sum to more than 1. Once this is not a distribution, it is also difficult to model it with standard autoregressive transformers.

Conclusion. When events are mutually exclusive in time, interventions on intensities act atomically as we expect: they do not shift probability mass to or from events that are not intervened. Interventions on competing event distributions coincide with intensity interventions only when the intervened event $a \in \mathcal{E}$ never wins a tie.

When concurrent events are allowed, intensities also require some convention on ordering since the events can be reordered within their concurrent time slot. Then intervening on the intensity

representation results in unintuitive effects such as creating correlations between events that are statistically independent. That is, unless we modify the definition of the survival function, and in turn change the normalization properties of $\lambda(t)$, which makes training a model based on such intensities difficult. For interventions on competing events, the position in the ordering determines whether the intervention creates an instantaneous effect (e.g. when it's first in the ordering and axiom C.1 is not satisfied) vs. not creating it (when it's last and the axiom is satisfied).

C.4 Transformations between representations

Since we may wish to be flexible about our assumptions on the instantaneous effects created by interventions, it is helpful to represent the same distribution under a different ordering. Given an ordering $\prec$, there is a simple transformation from any $P(T_{\text{next}}, E_{\text{next}})$ to a competing events representation under ordering $\prec$. We show it below.

Lemma C.7. *Let $P(T_{\text{next}}, E_{\text{next}})$ and an ordering $\prec$. Assume w.l.o.g that $e_1 \prec e_2 \prec \ldots \prec e_{|\mathcal{E}|}$. Ordering all tuples (t, e_j) by time t, and then by j within each time, let (t^*, j^*) be the first tuple for which $\sum_{t<t^*} \sum_j P(t, e_j) \sum_{j \leq j^*} P(t^*, e_j) = 1$.*

1. *There exists a competing events representation under ordering $\prec$ of $P(T_{\text{next}}, E_{\text{next}})$.*
2. *The representation $q_{e_j}(t)$ is unique for all entries where $t < t^*$ or $t = t^* \wedge j \leq j^*$.*

In short, we can represent any joint distribution, and the representation is unique up to some entries (t, e) where $P(t, e) = 0$. The proof is straightforward and given below, and the conclusion from its proof is that the procedure in algorithm 1 converts a joint probability table $P(E_{\text{next}}, T_{\text{next}})$ to a competing events representation. As explained in the main paper, this is useful for switching representations at inference time.

Proof. As explained in the likelihood of CEMs, eq. (2), we say $\{q_e\}_{e \in \mathcal{E}}$ is a representation of P when it satisfies for all $t < \infty$ and $e_j \in \mathcal{E}$:

$$P(T_{\text{next}} = t, E_{\text{next}} = e_j) = q_{e_j}(t) \prod_{j'=1}^{j-1} \left[1 - \sum_{k \leq t} q_{e_{j'}}(k)\right] \prod_{j'=j+1}^{|\mathcal{E}|} \left[1 - \sum_{k<t} q_{e_{j'}}(k)\right]. \tag{C.8}$$

Here we took by convention when summing or multiplying on empty sets that for $j = 1$ the item $\prod_{i=1}^{j-1} \left[1 - \sum_{k \leq t} q_{e_{j'}}(k)\right] := 1$; for $j = |\mathcal{E}|$ that $\prod_{i=j+1}^{|\mathcal{E}|} \left[1 - \sum_{k<t} q_{e_i}(k)\right] := 1$; for $t = 0$ that $\sum_{k<t} q_{e_i}(k) := 0$.
We will prove by induction on the order of $t = 0 \ldots T$ and within each t on the order of events $j = 1 \ldots |\mathcal{E}|$. Our claims are that setting

$$q_{e_j}(t) \leftarrow \frac{P(T_{\text{next}} = t, E_{\text{next}} = e_j)}{\prod_{i=1}^{j-1} \left[1 - \sum_{k \leq t} q_{e_i}(k)\right] \prod_{i=j+1}^{|\mathcal{E}|} \left[1 - \sum_{k<t} q_{e_i}(k)\right]}, \tag{C.9}$$

sequentially on the order we set above, then:

- *(a)* $1 - \sum_{k \leq t} q_{e_j}(k) > 0$ for all $t \leq t^*$ and $j < j^*$ and *(b)* $1 - \sum_{k<t} q_{e_j}(k) > 0$ for all $t < t^*$ and $j \geq j^*$.
- The obtained $\{q_e\}_{e \in \mathcal{E}}$ represents P when setting $q_e(\infty) \to 1 - \sum_{t=0}^{T} q_e(t)\ \forall e \in \mathcal{E}$, and it is unique up to t^*, j^* as described formally in our statement.

The second claim is what we stated in the lemma, and therefore proving it concludes the proof. It is easily obtained from the first claim, because under the first claim $\prod_{i=1}^{j-1} \left[1 - \sum_{k \leq t} q_{e_i}(k)\right] \prod_{i=j+1}^{|\mathcal{E}|} \left[1 - \sum_{k<t} q_{e_i}(k)\right] \neq 0$ for $t < t^*$ or $t = t^* \wedge j < j^*$, which means eq. (C.8) uniquely determines $q_{e_j}(t)$ to the value assigned by eq. (C.9). For coordinates t, j that are iterated after t^*, j^* we have $P(t, e_j) = 0$ and eq. (C.9) indeed assigns $q_{e_j}(t) \to 0$ that satisfies eq. (C.8). Yet, it is not a unique valid assignment

since the items $\prod_{j'=1}^{j-1}\left[1-\sum_{k\le t} q_{e_{j'}}(k)\right]\prod_{j'=j+1}^{|\mathcal{E}|}\left[1-\sum_{k<t} q_{e_{j'}}(k)\right]$ in eq. (C.8) may be 0 and allow other values of $q_{e_j}(t)$.

Hence, it is left to prove the first claim in the items above. For the base case where $t=0, j=1$: for condition *(a)*, if $j^*>1$ then $1-\sum_{k\le 0} q_{e_1}(k) = 1-P(0,e_1)>0$ and if $j^*=0$ then there are no j that satisfy $j<j^*$ and the statement is trivially true. For condition *(b)*, our conventions for empty sets give $\sum_{k<t} q_{e_j}(t) = \sum_{k<0} q_{e_1}(0) = 0$ which shows the condition holds.

For the induction step, assume that the statement holds for coordinates earlier than t, j. We will show that:

$$1-\sum_{k\le t} q_{e_j(t)} = \frac{1-\sum_{k<t}\sum_{j'} P(k,e_{j'}) - \sum_{j'\le j} P(t,e_{j'})}{\prod_{i=1}^{j-1}\left[1-\sum_{k\le t} q_{e_i}(k)\right]\prod_{i=j+1}^{|\mathcal{E}|}\left[1-\sum_{k<t} q_{e_i}(k)\right]} \tag{C.10}$$

Then the denominator is positive due to the inductive assumption, and the nominator is positive because (t,j) comes before (t^*,j^*) in our iterative order. Therefore the identity will complete our proof. To prove the identity we will once again use induction, where in the base case $t=0, j=1$ it reduces to $1-q_{e_1}(0) = 1-P(0,e_1)$ which holds trivially. For the step, we have

$$\begin{aligned}
1-\sum_{k\le t} q_{e_j(t)} &= 1-\sum_{k<t} q_{e_j(t)} - q_{e_j(t)} \\
&= 1-\sum_{k<t} q_{e_j(t)} - \frac{P(t,e_j)}{\prod_{i=1}^{j-1}\left[1-\sum_{k\le t} q_{e_i}(k)\right]\prod_{i=j+1}^{|\mathcal{E}|}\left[1-\sum_{k<t} q_{e_i}(k)\right]} \\
&= \frac{\prod_{i=1}^{j-1}\left[1-\sum_{k\le t} q_{e_i}(k)\right]\prod_{i=j}^{|\mathcal{E}|}\left[1-\sum_{k<t} q_{e_i}(k)\right] - P(t,e_j)}{\prod_{i=1}^{j-1}\left[1-\sum_{k\le t} q_{e_i}(k)\right]\prod_{i=j+1}^{|\mathcal{E}|}\left[1-\sum_{k<t} q_{e_i}(k)\right]} \\
&\overset{C.10 \text{ w. } j-1}{=} \frac{1-\sum_{k<t}\sum_{j'} P(k,e_{j'}) - \sum_{j'\le j} P(t,e_{j'})}{\prod_{i=1}^{j-1}\left[1-\sum_{k\le t} q_{e_i}(k)\right]\prod_{i=j+1}^{|\mathcal{E}|}\left[1-\sum_{k<t} q_{e_i}(k)\right]}.
\end{aligned}$$

Here, the last equality is from the induction hypothesis taken w.r.t coordinate $(t, j-1)$ (or $(t-1, |\mathcal{E}|)$ when $j=0$), where we multiplied eq. (C.10) by the denominator of the right-hand-side. This proves eq. (C.10) for all tuples that arrive before (t^*, j^*), and concludes our proof.

□

Algorithm 1 Conversion from joint distribution to CEM

Input: Joint distribution $P(T_{\text{next}}, E_{\text{next}})$, Ordering $e_1 \prec e_2 \prec \ldots \prec e_{|\mathcal{E}|}$
for $t = 0, \ldots, T$ **do**
  for $j = 1, \ldots, |\mathcal{E}|$ **do**
    $q_{e_j}(t) \leftarrow \frac{P(T_{\text{next}}=t, E_{\text{next}}=e_j)}{\prod_{i=1}^{j-1}\left[1-\sum_{k\le t} q_{e_i}(k)\right]\prod_{i=j+1}^{|\mathcal{E}|}\left[1-\sum_{k<t} q_{e_i}(k)\right]}$
for $e \in \mathcal{E}$ **do**
  $q_e(\infty) \leftarrow 1-\sum_{t=0}^{T} q_e(t)$

C.5 Identifiability assumptions

Our identifiability assumption from the main paper was as follows.

Assumption 4.1. *Let $f(\mathcal{H}) = f(\mathcal{H}^y)$ where $y \in \mathcal{E}$ is an outcome type. We assume the unobserved processes can be decomposed as $\mathcal{U} = [u_1, \ldots, u_l]$ such that for each u_i, where we denote $\mathcal{U}_{>i} = [u_{i+1}, \ldots u_l]$, it holds either that* (i) $q_a(T_a \mid \mathcal{H}_t, \mathcal{H}_t^{u_i}, \mathcal{H}_t^{\mathcal{U}_{>i}}) = q_a(T_a \mid \mathcal{H}_t, \mathcal{H}_t^{\mathcal{U}_{>i}})$, *or* (ii) $q_e(T_e \mid \mathcal{H}_t, \mathcal{H}_t^{u_i}, \mathcal{H}_t^{\mathcal{U}_{>i}}) = q_e(T_e \mid \mathcal{H}_t, \mathcal{H}_t^{\mathcal{U}_{>i}})$ *for all $e \in \{\mathcal{E} \setminus a\} \cup \mathcal{U}_{>i}$.*

This assumption boils down to eliminability (Røysland et al., 2022), in the special case where $\mathcal{N}_0$ in their notation is set to $\mathcal{E}$, i.e. all the observed events. This means we control for all the events that are recorded, and not only a subset of them. One could train a CEM only over a subset of the types $\mathcal{E}$, which induces a different identifiability assumption. Recall that when conditioning on $\mathcal{H}_t$ where $\max_{(t_k,e_k)\in\mathcal{H}_t} t_k < t$, it should be read as conditioning on $T_{\text{next}} \geq t$ as well. Written explicitly, $P(T_{\text{next}}, E_{\text{next}} \mid \mathcal{H}_t) = P(T_{\text{next}}, E_{\text{next}} \mid (t_1,e_1),\dots,(t_k,e_k), T_{\text{next}} \geq t)$. For a distribution over the next time and type $D(T_{\text{next}}, E_{\text{next}})$ supported on $\overline{\mathcal{T}} \times (\overline{\mathcal{E}} \cup \mathcal{U})$, we denote the marginal distribution over $\overline{\mathcal{T}} \times \overline{\mathcal{E}}$ by $D|_{\mathcal{E}}$. That is, we marginalize over all the unobservable events that can happen between time t and the next observable event $(t_{\text{next}}, e_{\text{next}})$.

$$D|_{\mathcal{E}}(T_{\text{next}} = t_{\text{next}}, E_{\text{next}} = e_{\text{next}} \mid \mathcal{H}_t) := \sum_{k=0}^{\infty} \sum_{(\boldsymbol{t}^u, \boldsymbol{e}^u)\in(\mathcal{U}\times\mathcal{T})^k} D(T_{\text{next}} = t, E_{\text{next}} = e \mid \mathcal{H}_{t_k^u}, \mathcal{H}^{\mathcal{U}}_{t_k^u}) \prod_{j=0}^{k-1} D(T_{\text{next}} = t^u_{j+1}, E_{\text{next}} = e^u_{j+1} \mid \mathcal{H}_{t_j^u}, \mathcal{H}^{\mathcal{U}}_{t_j^u}).$$

A competing events representation of such a joint distribution is consisted of distributions $\{q|_{\mathcal{E},e}(T_e \mid \mathcal{H}_t)\}_{e\in\mathcal{E}}$. It is guaranteed to exist due to lemma C.7, and it is also easy to see that if $q_{e'}(T_{e'} \mid \mathcal{H}_t, \mathcal{H}^{\mathcal{U}}_t) = q_{e'}(T_{e'} \mid \mathcal{H}_t)$ in the full distribution D then setting $q|_{\mathcal{E},e'} := q_{e'}$ in a competing events representation $\{q|_{\mathcal{E},e}(T_e \mid \mathcal{H}_t)\}_{e\in\mathcal{E}}$ of $D|_{\mathcal{E}}$ still gives a competing events representation of $D|_{\mathcal{E}}$.

With this notation, the observational distribution is $P|_{\mathcal{E}}$, whereas P is the complete distribution that includes unobserved event types $\mathcal{U}$ as well. The interventional distribution is $\tilde{P}$, while the one that we use for estimation is $\widetilde{P|_{\mathcal{E}}}$. The claim is then as follows.

Lemma C.8. *Under assumption 4.1 and overlap ($\tilde{P} \ll P$), we have* $\mathbb{E}_{\tilde{P}}[f(\mathcal{H}^y) \mid \mathcal{H}_t] = \mathbb{E}_{\widetilde{P|_{\mathcal{E}}}}[f(\mathcal{H}^y) \mid \mathcal{H}_t]$.

Proof. **Note about ties:** For simplicity, we omit the ordering $\prec$ since including it results in additional terms where the time argument changes between t' and $t'+1$ for types that arrive before or after an event e. Instead we just write t' for all survival terms in the likelihood, and items where this should be replaced with $t'+1$ should be understood from context. We assume that the causal model P uses some ordering $\prec$, and the intervention in $\tilde{P}$ respects that ordering (i.e. only q_a is replaced, and not the tie breaking mechanism). When we marginalize over latent variables, the ordering among observed events is the one induced by the full ordering $\prec$.

The estimand $\mathbb{E}_{\tilde{P}}[f(\mathcal{H}^y) \mid \mathcal{H}_t]$ can be broken down recursively using the law of total expectation, until the next observed event as follows (notice that the last event in a sequence is always an observed one, so this decomposition is valid):

$$\mathop{\mathbb{E}}_{\tilde{P}}[f(\mathcal{H}^y) \mid \mathcal{H}_t] = \sum_{(t',e')\in\overline{\mathcal{E}}\times\overline{\mathcal{T}}} \sum_{k=0}^{\infty} \sum_{(\boldsymbol{t}^u,\boldsymbol{e}^u)\in(\mathcal{U}\times\mathcal{T})^k} \tilde{P}(T_{\text{next}} = t', E_{\text{next}} = e' \mid \mathcal{H}_{t_k^u}, \mathcal{H}^{\mathcal{U}}_{t_k^u}) \prod_{j=0}^{k-1} \tilde{P}(T_{\text{next}} = t^u_{j+1}, E_{\text{next}} = e^u_{j+1} \mid \mathcal{H}_{t_j^u}, \mathcal{H}^{\mathcal{U}}_{t_j^u}) \mathop{\mathbb{E}}_{\tilde{P}}[f(\mathcal{H}^y) \mid \mathcal{H}_{t'}, \mathcal{H}^{\mathcal{U}}_{t'}] \quad \text{(C.11)}$$

Here k iterates over the possible number of unobserved events between time t and the next obeserved event, and the inner sum on $\boldsymbol{t}^u, \boldsymbol{e}^u$ iterates over all possible times and types for these unobserved events. To keep notation light, we leave implicit the average over the unobserved history before t, with weights $P(\mathcal{H}^{\mathcal{U}}_t \mid \mathcal{H}_t) = \tilde{P}(\mathcal{H}^{\mathcal{U}}_t \mid \mathcal{H}_t)$, which coincide since the intervention acts only after t. The proof proceeds by induction on the number of unobserved types, and within that an induction on the number of observed events in the sequence.

One unobserved type. Assume that u_1 is the only unobserved event type.

Case 1: If it satisfies (i) $q_a(T_a \mid \mathcal{H}_t, \mathcal{H}^{\mathcal{U}}_t) = q_a(T_a \mid \mathcal{H}_t, \mathcal{H}^{u_1}_t) = q_a(T_a \mid \mathcal{H}_t)$ then for any $t' \in \mathcal{T}, e' \in \mathcal{U}\cup\mathcal{E}\setminus a, j \geq 0$,

$$\tilde{P}(T_{\text{next}} = t', E_{\text{next}} = e' \mid \mathcal{H}_{t_j^u}, \mathcal{H}^{\mathcal{U}}_{t_j^u}) = \tilde{q}_{e'}(t' \mid \mathcal{H}_{t_j^u}, \mathcal{H}^{\mathcal{U}}_{t_j^u}) \prod_{e\neq e'} 1 - \tilde{F}_e[t' \mid \mathcal{H}_{t_j^u}, \mathcal{H}^{\mathcal{U}}_{t_j^u}] =$$

$$q_{e'}(t' \mid \mathcal{H}_{t_j^u}, \mathcal{H}^{\mathcal{U}}_{t_j^u})(1-\tilde{F}_a[t' \mid \mathcal{H}_{t_j^u}, \mathcal{H}^{\mathcal{U}}_{t_j^u}]) \prod_{e \notin \{e', a\}} 1 - F_e[t' \mid \mathcal{H}_{t_j^u}, \mathcal{H}^{\mathcal{U}}_{t_j^u}] =$$

$$q_{e'}(t' \mid \mathcal{H}_{t_j^u}, \mathcal{H}^{\mathcal{U}}_{t_j^u})(1-\tilde{F}_a[t' \mid \mathcal{H}_{t_j^u}]) \frac{1 - F_a[t' \mid \mathcal{H}_{t_j^u}, \mathcal{H}^{\mathcal{U}}_{t_j^u}]}{1 - F_a[t' \mid \mathcal{H}_{t_j^u}]} \prod_{e \notin \{e', a\}} 1 - F_e[t' \mid \mathcal{H}_{t_j^u}, \mathcal{H}^{\mathcal{U}}_{t_j^u}] =$$

$$P(T_{\text{next}} = t', E_{\text{next}} = e' \mid \mathcal{H}_{t_j^u}, \mathcal{H}^{\mathcal{U}}_{t_j^u}) \frac{1 - \tilde{F}_a[t' \mid \mathcal{H}_{t_j^u}]}{1 - F_a[t' \mid \mathcal{H}_{t_j^u}]}.$$

Here, we used the convention $t_0^u = t$. Similarly for $e' = a$ we'll have $\tilde{P}(T_{\text{next}} = t', E_{\text{next}} = a \mid \mathcal{H}_{t_j^u}, \mathcal{H}^{\mathcal{U}}_{t_j^u}) = P(T_{\text{next}} = t', E_{\text{next}} = a \mid \mathcal{H}_{t_j^u}, \mathcal{H}^{\mathcal{U}}_{t_j^u}) \frac{\tilde{q}_a(t' \mid \mathcal{H}_{t_j^u})}{q_a(t' \mid \mathcal{H}_{t_j^u})}$. Now, letting $\{q|_{\mathcal{E},e}\}$ be a competing events representation of $P|_{\mathcal{E}}$, a product of survival terms ratios like those that appear above is:

$$\prod_{j=0}^{k-1} \frac{1 - \tilde{F}_a[t^u_{j+1} \mid \mathcal{H}_{t_j^u}]}{1 - F_a[t^u_{j+1} \mid \mathcal{H}_{t_j^u}]} = \prod_{j=0}^{k-1} \frac{1 - \tilde{F}_a[t^u_{j+1} \mid \mathcal{H}_{t_j^u}] \prod_{e \in \mathcal{E} \setminus a} 1 - F|_{\mathcal{E},e}[t^u_{j+1} \mid \mathcal{H}_{t_j^u}]}{1 - F_a[t^u_{j+1} \mid \mathcal{H}_{t_j^u}] \prod_{e \in \mathcal{E} \setminus a} 1 - F|_{\mathcal{E},e}[t^u_{j+1} \mid \mathcal{H}_{t_j^u}]} =$$

$$\prod_{j=0}^{k-1} \frac{\widetilde{P|_{\mathcal{E}}}(T_{\text{next}} > t^u_{j+1} \mid \mathcal{H}_{t_j^u})}{P|_{\mathcal{E}}(T_{\text{next}} > t^u_{j+1} \mid \mathcal{H}_{t_j^u})} =$$

$$\frac{\widetilde{P|_{\mathcal{E}}}(T_{\text{next}} > t^u_k \mid \mathcal{H}_t)}{P|_{\mathcal{E}}(T_{\text{next}} > t^u_k \mid \mathcal{H}_t)} \tag{C.12}$$

The first step is valid because we multiply the numerator and denominator by the same non-zero items (as we assume an unobserved event has a positive probability of occurring at t_j^u for each $j = 1, \dots k$), which means we multiply the entire expression by 1. The second transition is from the definition of $P|_{\mathcal{E}}$ since $\{q|_{\mathcal{E},e}\}_{e \in \mathcal{E}}$ is a representation of $P|_{\mathcal{E}}$ and q_a is independent of the unobserved history, setting $q|_{\mathcal{E},a} := q_a$ still represents $P|_{\mathcal{E}}$. This explains the denominator of the expression, while the numerator is correct since it replaces the distribution $q|_{\mathcal{E},a}$ with $\tilde{q}_a$ which forms the distribution used in estimation. Finally, the third step is due to the chain rule.

Similarly to eq. (C.12), for the last event (i.e. the observed one, conditioned on all the k unobserved events), if $e' \neq a$ then:

$$\frac{1 - \tilde{F}_a[t' \mid \mathcal{H}_{t_k^u}]}{1 - F_a[t' \mid \mathcal{H}_{t_k^u}]} = \frac{q|_{\mathcal{E},e'}(t' \mid \mathcal{H}_{t_k^u}) \cdot (1 - \tilde{F}_a[t' \mid \mathcal{H}_{t_k^u}]) \prod_{e \in \mathcal{E} \setminus \{e', a\}} 1 - F|_{\mathcal{E},e}[t' \mid \mathcal{H}_{t_k^u}]}{q|_{\mathcal{E},e'}(t' \mid \mathcal{H}_{t_k^u}) \cdot (1 - F_a[t' \mid \mathcal{H}_{t_k^u}]) \prod_{e \in \mathcal{E} \setminus \{e', a\}} 1 - F|_{\mathcal{E},e}[t' \mid \mathcal{H}_{t_k^u}]} =$$

$$\frac{\widetilde{P|_{\mathcal{E}}}(T_{\text{next}} = t', E_{\text{next}} = e' \mid \mathcal{H}_{t_k^u})}{P|_{\mathcal{E}}(T_{\text{next}} = t', E_{\text{next}} = e' \mid \mathcal{H}_{t_k^u})}.$$

If $e' = a$ then similarly we will have $\frac{\tilde{q}_a(t' \mid \mathcal{H}_{t_k^u})}{q_a(t' \mid \mathcal{H}_{t_k^u})} = \frac{\widetilde{P|_{\mathcal{E}}}(T_{\text{next}}=t', E_{\text{next}}=e' \mid \mathcal{H}_{t_k^u})}{P|_{\mathcal{E}}(T_{\text{next}}=t', E_{\text{next}}=e' \mid \mathcal{H}_{t_k^u})}$. Overall in both cases, this lets us conclude that

$$\tilde{P}(T_{\text{next}} = t', E_{\text{next}} = e' \mid \mathcal{H}_{t_k^u}, \mathcal{H}^{\mathcal{U}}_{t_k^u}) =$$

$$P(T_{\text{next}} = t', E_{\text{next}} = e' \mid \mathcal{H}_{t_k^u}, \mathcal{H}^{\mathcal{U}}_{t_k^u}) \frac{\widetilde{P|_{\mathcal{E}}}(T_{\text{next}} = t', E_{\text{next}} = e' \mid \mathcal{H}_{t_k^u})}{P|_{\mathcal{E}}(T_{\text{next}} = t', E_{\text{next}} = e' \mid \mathcal{H}_{t_k^u})}. \tag{C.13}$$

Plugging this into eq. (C.11), we get:

$$\mathop{\mathbb{E}}_{\tilde{P}}[f(\mathcal{H}^y) \mid \mathcal{H}_t] = \sum_{(t', e') \in \overline{\mathcal{T}} \times \overline{\mathcal{E}}} \sum_{k=0}^{\infty} \sum_{(\boldsymbol{t}^u, \boldsymbol{e}^u) \in (\mathcal{T} \times \mathcal{U})^k} \tilde{P}(T_{\text{next}} = t', E_{\text{next}} = e' \mid \mathcal{H}_{t_k^u}, \mathcal{H}^{\mathcal{U}}_{t_k^u}) \cdot$$

$$\prod_{j=0}^{k-1} \frac{1 - \tilde{F}_a[t^u_{j+1} \mid \mathcal{H}_{t_j^u}]}{1 - F_a[t^u_{j+1} \mid \mathcal{H}_{t_j^u}]} P(T_{\text{next}} = t^u_{j+1}, E_{\text{next}} = e^u_{j+1} \mid \mathcal{H}_{t_j^u}, \mathcal{H}^{\mathcal{U}}_{t_j^u}) \mathop{\mathbb{E}}_{\tilde{P}}[f(\mathcal{H}^y) \mid \mathcal{H}_{t'}, \mathcal{H}^{\mathcal{U}}_{t'}] \overset{C.12}{=}$$

$$\sum_{(t', e') \in \overline{\mathcal{T}} \times \overline{\mathcal{E}}} \sum_{k=0}^{\infty} \sum_{(\boldsymbol{t}^u, \boldsymbol{e}^u) \in (\mathcal{T} \times \mathcal{U})^k} \tilde{P}(T_{\text{next}} = t', E_{\text{next}} = e' \mid \mathcal{H}_{t_k^u}, \mathcal{H}^{\mathcal{U}}_{t_k^u}) \cdot \frac{\widetilde{P|_{\mathcal{E}}}(T_{\text{next}} > t^u_k \mid \mathcal{H}_t)}{P|_{\mathcal{E}}(T_{\text{next}} > t^u_k \mid \mathcal{H}_t)}.$$

$$\prod_{j=0}^{k-1} P(T_{\text{next}} = t^u_{j+1}, E_{\text{next}} = e^u_{j+1} \mid \mathcal{H}_{t^u_j}, \mathcal{H}^{\mathcal{U}}_{t^u_j}) \mathop{\mathbb{E}}_{\tilde{P}}[f(\mathcal{H}^y) \mid \mathcal{H}_{t'}, \mathcal{H}^{\mathcal{U}}_{t'}] \overset{C.13}{=}$$

$$\sum_{(t',e')\in\overline{\mathcal{T}}\times\overline{\mathcal{E}}} \sum_{k=0}^{\infty} \sum_{(\boldsymbol{t}^u,\boldsymbol{e}^u)\in(\mathcal{T}\times\mathcal{U})^k} \frac{\widetilde{P|_{\mathcal{E}}}(t',e' \mid \mathcal{H}_{t^u_k})}{P|_{\mathcal{E}}(t',e' \mid \mathcal{H}_{t^u_k})} P(T_{\text{next}} = t', E_{\text{next}} = e' \mid \mathcal{H}_{t^u_k}, \mathcal{H}^{\mathcal{U}}_{t^u_k}) \cdot$$

$$\frac{\widetilde{P|_{\mathcal{E}}}(T_{\text{next}} > t^u_k \mid \mathcal{H}_t)}{P|_{\mathcal{E}}(T_{\text{next}} > t^u_k \mid \mathcal{H}_t)} \prod_{j=0}^{k-1} P(T_{\text{next}} = t^u_{j+1}, E_{\text{next}} = e^u_{j+1} \mid \mathcal{H}_{t^u_j}, \mathcal{H}^{\mathcal{U}}_{t^u_j}) \mathop{\mathbb{E}}_{\tilde{P}}[f(\mathcal{H}^y) \mid \mathcal{H}_{t'}, \mathcal{H}^{\mathcal{U}}_{t'}] =$$

$$\sum_{(t',e')\in\overline{\mathcal{T}}\times\overline{\mathcal{E}}} \frac{\widetilde{P|_{\mathcal{E}}}(t',e' \mid \mathcal{H}_t)}{P|_{\mathcal{E}}(t',e' \mid \mathcal{H}_t)} \sum_{k=0}^{\infty} \sum_{(\boldsymbol{t}^u,\boldsymbol{e}^u)\in(\mathcal{T}\times\mathcal{U})^k} P(T_{\text{next}} = t', E_{\text{next}} = e' \mid \mathcal{H}_{t^u_k}, \mathcal{H}^{\mathcal{U}}_{t^u_k}) \cdot$$

$$\prod_{j=0}^{k-1} P(T_{\text{next}} = t^u_{j+1}, E_{\text{next}} = e^u_{j+1} \mid \mathcal{H}_{t^u_j}, \mathcal{H}^{\mathcal{U}}_{t^u_j}) \mathop{\mathbb{E}}_{\tilde{P}}[f(\mathcal{H}^y) \mid \mathcal{H}_{t'}, \mathcal{H}^{\mathcal{U}}_{t'}] =$$

$$\sum_{(t',e')\in\overline{\mathcal{T}}\times\overline{\mathcal{E}}} \frac{\widetilde{P|_{\mathcal{E}}}(t',e' \mid \mathcal{H}_t)}{P|_{\mathcal{E}}(t',e' \mid \mathcal{H}_t)} P|_{\mathcal{E}}(T_{\text{next}} = t', E_{\text{next}} = e' \mid \mathcal{H}_t) \mathop{\mathbb{E}}_{\tilde{P}}[f(\mathcal{H}^y) \mid \mathcal{H}_{t'}] =$$

$$\mathop{\mathbb{E}}_{t',e'\sim\widetilde{P|_{\mathcal{E}}}(T_{\text{next}},E_{\text{next}} \mid \mathcal{H}_t)} \mathop{\mathbb{E}}_{\tilde{P}}[f(\mathcal{H}^y) \mid \mathcal{H}_{t'}]$$

The second-to-last equality uses two facts. First, after averaging over the unobserved history before t, the sum over unobserved paths of the probabilities under P is the probability of the next observed event under P, i.e. $P|_{\mathcal{E}}(t', e' \mid \mathcal{H}_t)$ times the posterior $P(\mathcal{H}^{\mathcal{U}}_{t'} \mid \mathcal{H}_{t'})$ of the unobserved history. Second, these posterior weights are taken under P, while the inner expectation is taken under $\tilde{P}$, so the law of total expectation requires $P(\mathcal{H}^{\mathcal{U}}_{t'} \mid \mathcal{H}_{t'}) = \tilde{P}(\mathcal{H}^{\mathcal{U}}_{t'} \mid \mathcal{H}_{t'})$. This holds by condition (i): by eqs. (C.12) and (C.13), the ratio between the probabilities that $\tilde{P}$ and P assign to $(\mathcal{H}_{t'}, \mathcal{H}^{\mathcal{U}}_{t'})$ given $(\mathcal{H}_t, \mathcal{H}^{\mathcal{U}}_t)$ depends only on the observed history, and the posteriors before t coincide, so the ratio cancels when normalizing over $\mathcal{H}^{\mathcal{U}}_{t'}$.

Applying this identity recursively, we arrive at the desired identity $\mathbb{E}_{\tilde{P}}[f(\mathcal{H}^y) \mid \mathcal{H}_t] = \mathbb{E}_{\mathcal{H}\sim\widetilde{P|_{\mathcal{E}}}}[f(\mathcal{H}^y) \mid \mathcal{H}_t]$

Case 2: Suppose u_1 satisfies (ii), $q_e(T_e \mid \mathcal{H}_s, \mathcal{H}^{\mathcal{U}}_s) = q_e(T_e \mid \mathcal{H}_s, \mathcal{H}^{u_1}_s) = q_e(T_e \mid \mathcal{H}_s)$ for all $e \in \mathcal{E} \setminus a$ and all s. Let $\{q_e\}_{e\in\mathcal{E}\cup\mathcal{U}}$ be a competing events representation of the observational P, so that $\{q_e\}_{e\in(\mathcal{E}\setminus a)\cup\mathcal{U}} \cup \{\tilde{q}_a\}$ is a competing events representation of $\tilde{P}$ after time t. The key property is that under $\tilde{P}$, after time t, no clock of a type in $\mathcal{E}$ depends on the unobserved history: for $e \neq a$ this is condition (ii), and for a it holds because $\tilde{q}_a$ depends only on observed events. We will show that

- The expectation $\mathbb{E}_{\tilde{P}}[f(\mathcal{H}^y) \mid \mathcal{H}_t]$ can be written as a function of observed clocks and histories alone as follows:

$$\mathop{\mathbb{E}}_{\tilde{P}}[f(\mathcal{H}^y) \mid \mathcal{H}_t] = \sum_{(t',e')\in\mathcal{T}\times\mathcal{E}} \left(\tilde{q}_{e'}(t' \mid \mathcal{H}_t) \prod_{e\in\mathcal{E}\setminus e'} 1 - \tilde{F}_{e'}[t' \mid \mathcal{H}_t] \right) \cdot \mathop{\mathbb{E}}_{\tilde{P}}[f(\mathcal{H}^y) \mid \mathcal{H}_{t'}] +$$
$$\prod_{e\in\mathcal{E}} \tilde{q}_{e'}(\infty \mid \mathcal{H}_t) \mathop{\mathbb{E}}_{\tilde{P}}[f(\mathcal{H}^y) \mid \mathcal{H}_t \cup (\infty, \text{END})]. \tag{C.14}$$

  Note that here $\tilde{q}_e = q_e$ for all $e \in \mathcal{E} \setminus a$ and it only changes q_a from the observational P.

- $\{q_e\}_{e\in\mathcal{E}\setminus a} \cup \{\tilde{q}_a\}$ is a representation of $\widetilde{P|_{\mathcal{E}}}$. Therefore,

$$\tilde{q}_{e'}(t' \mid \mathcal{H}_t) \prod_{e\in\mathcal{E}\setminus e'} 1 - \tilde{F}_{e'}[t' \mid \mathcal{H}_t] = \widetilde{P|_{\mathcal{E}}}(T_{\text{next}} = t', E_{\text{next}} = e' \mid \mathcal{H}_t),$$
$$\prod_{e\in\mathcal{E}} \tilde{q}_{e'}(\infty \mid \mathcal{H}_t) = \widetilde{P|_{\mathcal{E}}}(T_{\text{next}} = \infty, E_{\text{next}} = \text{END} \mid \mathcal{H}_t).$$

Then plugging the equations from the second bullet into the identity in the first gives $\mathbb{E}_{\tilde{P}}[f(\mathcal{H}^y) \mid \mathcal{H}_t] = \mathbb{E}_{t',e' \sim \widetilde{P|_{\mathcal{E}}}(T_{\text{next}}, E_{\text{next}} \mid \mathcal{H}_t)} \mathbb{E}_{\tilde{P}}[f(\mathcal{H}^y) \mid \mathcal{H}_{t'}]$, and applying it recursively as in case 1 gives $\mathbb{E}_{\tilde{P}}[f(\mathcal{H}^y) \mid \mathcal{H}_t] = \mathbb{E}_{\widetilde{P|_{\mathcal{E}}}}[f(\mathcal{H}^y) \mid \mathcal{H}_t]$ as we claimed. It remains to prove the two bullets.

The next observed event under $\tilde{P}$. Fix $s \geq t$ and histories $\mathcal{H}_s, \mathcal{H}^{\mathcal{U}}_s$, and let $(T^{\mathcal{E}}_{\text{next}}, E^{\mathcal{E}}_{\text{next}})$ denote the next event of a type in $\mathcal{E}$. Expanding over the events of u_1 that precede it, as in eq. (C.11) with $t^u_0 = s$,

$$\tilde{P}(T^{\mathcal{E}}_{\text{next}} = t', E^{\mathcal{E}}_{\text{next}} = e' \mid \mathcal{H}_s, \mathcal{H}^{\mathcal{U}}_s) = \sum_{k=0}^{\infty} \sum_{\boldsymbol{t}^u \in \mathcal{T}^k} \tilde{P}(T_{\text{next}} = t', E_{\text{next}} = e' \mid \mathcal{H}_{t^u_k}, \mathcal{H}^{\mathcal{U}}_{t^u_k}) \prod_{j=0}^{k-1} \tilde{P}(T_{\text{next}} = t^u_{j+1}, E_{\text{next}} = u_1 \mid \mathcal{H}_{t^u_j}, \mathcal{H}^{\mathcal{U}}_{t^u_j}),$$

where terms with $t^u_k > t'$ vanish. Under $\tilde{P}$, the clocks of $\mathcal{E}$ do not depend on $\mathcal{H}^{\mathcal{U}}$, so

$$\tilde{P}(T_{\text{next}} = t^u_{j+1}, E_{\text{next}} = u_1 \mid \mathcal{H}_{t^u_j}, \mathcal{H}^{\mathcal{U}}_{t^u_j}) = q_{u_1}(t^u_{j+1} \mid \mathcal{H}_{t^u_j}, \mathcal{H}^{\mathcal{U}}_{t^u_j}) \prod_{e \in \mathcal{E}} 1 - \tilde{F}_e[t^u_{j+1} \mid \mathcal{H}_{t^u_j}],$$

$$\tilde{P}(T_{\text{next}} = t', E_{\text{next}} = e' \mid \mathcal{H}_{t^u_k}, \mathcal{H}^{\mathcal{U}}_{t^u_k}) = \tilde{q}_{e'}(t' \mid \mathcal{H}_{t^u_k}) \big(1 - F_{u_1}[t' \mid \mathcal{H}_{t^u_k}, \mathcal{H}^{\mathcal{U}}_{t^u_k}]\big) \prod_{e \in \mathcal{E} \setminus e'} 1 - \tilde{F}_e[t' \mid \mathcal{H}_{t^u_k}].$$

Since there are no observed events in $[s, t^u_k]$, conditioning the clock of $e \in \mathcal{E}$ on $\mathcal{H}_{t^u_j}$ instead of $\mathcal{H}_s$ only adds that e has not occurred by t^u_j. By the chain rule,

$$1 - \tilde{F}_e[t'' \mid \mathcal{H}_{t^u_j}] = \frac{1 - \tilde{F}_e[t'' \mid \mathcal{H}_s]}{1 - \tilde{F}_e[t^u_j \mid \mathcal{H}_s]}, \qquad \tilde{q}_e(t'' \mid \mathcal{H}_{t^u_j}) = \frac{\tilde{q}_e(t'' \mid \mathcal{H}_s)}{1 - \tilde{F}_e[t^u_j \mid \mathcal{H}_s]},$$

where survival terms are read as in the note about ties, so that $1 - \tilde{F}_e[s \mid \mathcal{H}_s] = 1$. The terms of each $e \in \mathcal{E} \setminus e'$ then telescope,

$$\prod_{j=0}^{k-1} \frac{1 - \tilde{F}_e[t^u_{j+1} \mid \mathcal{H}_s]}{1 - \tilde{F}_e[t^u_j \mid \mathcal{H}_s]} \cdot \frac{1 - \tilde{F}_e[t' \mid \mathcal{H}_s]}{1 - \tilde{F}_e[t^u_k \mid \mathcal{H}_s]} = 1 - \tilde{F}_e[t' \mid \mathcal{H}_s],$$

and the terms of e' telescope in the same way to $\tilde{q}_{e'}(t' \mid \mathcal{H}_s)$. Therefore

$$\begin{aligned}\tilde{P}(T^{\mathcal{E}}_{\text{next}} = t', E^{\mathcal{E}}_{\text{next}} = e' \mid \mathcal{H}_s, \mathcal{H}^{\mathcal{U}}_s) =& \tilde{q}_{e'}(t' \mid \mathcal{H}_s) \prod_{e \in \mathcal{E} \setminus e'} 1 - \tilde{F}_e[t' \mid \mathcal{H}_s] \cdot \\ & \sum_{k=0}^{\infty} \sum_{\boldsymbol{t}^u \in \mathcal{T}^k} \big(1 - F_{u_1}[t' \mid \mathcal{H}_{t^u_k}, \mathcal{H}^{\mathcal{U}}_{t^u_k}]\big) \prod_{j=0}^{k-1} q_{u_1}(t^u_{j+1} \mid \mathcal{H}_{t^u_j}, \mathcal{H}^{\mathcal{U}}_{t^u_j}) \\ =& \tilde{q}_{e'}(t' \mid \mathcal{H}_s) \prod_{e \in \mathcal{E} \setminus e'} 1 - \tilde{F}_e[t' \mid \mathcal{H}_s]. \end{aligned} \tag{C.15}$$

The last equality holds because the sum is the total probability, over all numbers k and times $\boldsymbol{t}^u$, that the events of u_1 before t' occur exactly at $\boldsymbol{t}^u$ in a model where u_1 is the only event type, which equals 1. The same calculation gives the probability of END. In particular, the right-hand side does not depend on $\mathcal{H}^{\mathcal{U}}_s$.

Proof of the first bullet. We show by induction on n that under $\tilde{P}$, the probability of the next n observed events given $(\mathcal{H}_s, \mathcal{H}^{\mathcal{U}}_s)$ does not depend on $\mathcal{H}^{\mathcal{U}}_s$. For $n = 1$ this is eq. (C.15). For $n > 1$, condition on the first observed event (t', e') and sum over the unobserved history $\mathcal{H}^{\mathcal{U}}_{t'}$ up to it. By the induction hypothesis, the probability of the remaining $n - 1$ observed events given $(\mathcal{H}_{t'}, \mathcal{H}^{\mathcal{U}}_{t'})$ is a function of $\mathcal{H}_{t'}$ only, so it can be taken out of the sum, and what remains is the probability of (t', e') given $(\mathcal{H}_s, \mathcal{H}^{\mathcal{U}}_s)$, which does not depend on $\mathcal{H}^{\mathcal{U}}_s$ by eq. (C.15). Consequently, the law of the observed trajectory after s under $\tilde{P}$ depends only on $\mathcal{H}_s$, and since $f(\mathcal{H}^y)$ is a function of the observed trajectory, $\mathbb{E}_{\tilde{P}}[f(\mathcal{H}^y) \mid \mathcal{H}_s, \mathcal{H}^{\mathcal{U}}_s] = \mathbb{E}_{\tilde{P}}[f(\mathcal{H}^y) \mid \mathcal{H}_s]$. Applying the law of total expectation

to the next observed event and using eq. (C.15) at $s = t$ gives eq. (C.14).

Proof of the second bullet. Under the observational P, the clocks of $e \in \mathcal{E} \setminus a$ do not depend on $\mathcal{H}^{\mathcal{U}}$, but the clock of a may. Repeating the calculation that led to eq. (C.15) under P, the terms of $e \in \mathcal{E} \setminus a$ telescope as before, while the terms of a remain inside the sum over the events of u_1:

$$P(T^{\mathcal{E}}_{\text{next}} = t', E^{\mathcal{E}}_{\text{next}} = e' \mid \mathcal{H}_s, \mathcal{H}^{\mathcal{U}}_s) = q_{e'}(t' \mid \mathcal{H}_s)\big(1 - G_a[t' \mid \mathcal{H}_s, \mathcal{H}^{\mathcal{U}}_s]\big) \cdot \prod_{e \in \mathcal{E} \setminus \{e', a\}} 1 - F_e[t' \mid \mathcal{H}_s], \quad \forall e' \in \mathcal{E} \setminus a,$$

$$P(T^{\mathcal{E}}_{\text{next}} = t', E^{\mathcal{E}}_{\text{next}} = a \mid \mathcal{H}_s, \mathcal{H}^{\mathcal{U}}_s) = g_a(t' \mid \mathcal{H}_s, \mathcal{H}^{\mathcal{U}}_s) \prod_{e \in \mathcal{E} \setminus a} 1 - F_e[t' \mid \mathcal{H}_s],$$

where

$$g_a(t' \mid \mathcal{H}_s, \mathcal{H}^{\mathcal{U}}_s) = \sum_{k=0}^{\infty} \sum_{\boldsymbol{t}^u \in \mathcal{T}^k} q_a(t' \mid \mathcal{H}_{t^u_k}, \mathcal{H}^{\mathcal{U}}_{t^u_k})\big(1 - F_{u_1}[t' \mid \mathcal{H}_{t^u_k}, \mathcal{H}^{\mathcal{U}}_{t^u_k}]\big) \prod_{j=0}^{k-1} q_{u_1}(t^u_{j+1} \mid \mathcal{H}_{t^u_j}, \mathcal{H}^{\mathcal{U}}_{t^u_j})\big(1 - F_a[t^u_{j+1} \mid \mathcal{H}_{t^u_j}, \mathcal{H}^{\mathcal{U}}_{t^u_j}]\big)$$

is the probability that the first event of a occurs at t' in a model with event types $\{a, u_1\}$ only, and G_a is the corresponding CDF. The other terms do not depend on $\mathcal{H}^{\mathcal{U}}_s$, so averaging over $P(\mathcal{H}^{\mathcal{U}}_s \mid \mathcal{H}_s)$ only affects g_a and G_a. Defining $q|_{\mathcal{E},a}(t' \mid \mathcal{H}_s) := \sum_{\mathcal{H}^{\mathcal{U}}_s} P(\mathcal{H}^{\mathcal{U}}_s \mid \mathcal{H}_s)\, g_a(t' \mid \mathcal{H}_s, \mathcal{H}^{\mathcal{U}}_s)$, whose CDF $F|_{\mathcal{E},a}$ is the corresponding average of G_a, we obtain

$$P|_{\mathcal{E}}(T_{\text{next}} = t', E_{\text{next}} = e' \mid \mathcal{H}_s) = q_{e'}(t' \mid \mathcal{H}_s)\,\big(1 - F|_{\mathcal{E},a}[t' \mid \mathcal{H}_s]\big) \cdot \prod_{e \in \mathcal{E} \setminus \{e', a\}} 1 - F_e[t' \mid \mathcal{H}_s], \quad e' \neq a,$$

$$P|_{\mathcal{E}}(T_{\text{next}} = t', E_{\text{next}} = a \mid \mathcal{H}_s) = q|_{\mathcal{E},a}(t' \mid \mathcal{H}_s) \prod_{e \in \mathcal{E} \setminus a} 1 - F_e[t' \mid \mathcal{H}_s].$$

Hence $\{q_e\}_{e \in \mathcal{E} \setminus a} \cup \{q|_{\mathcal{E},a}\}$ is a competing events representation of $P|_{\mathcal{E}}$. By lemma C.7, the representation under $\prec$ is unique on histories of positive probability, so $\widetilde{P|_{\mathcal{E}}}$, obtained by replacing the clock of a with $\tilde{q}_a$, is represented by $\{q_e\}_{e \in \mathcal{E} \setminus a} \cup \{\tilde{q}_a\}$. This gives the second bullet.

Multiple unobserved types. The proof for a single unobserved type does not use that the types in $\mathcal{E}$ are observed. It only uses that a, y and the types that $\tilde{q}_a$ depends on are in $\mathcal{E}$. Moreover, it holds for every measurable function f of the trajectory of $\mathcal{E}$, and both $\tilde{P}|_{\mathcal{E}}$ and $\widetilde{P|_{\mathcal{E}}}$ coincide with $P|_{\mathcal{E}}$ before t. Hence it shows the equality of distributions $\tilde{P}|_{\mathcal{E}} = \widetilde{P|_{\mathcal{E}}}$ for any competing events model over $\mathcal{E} \cup \{u\}$ in which u satisfies (i) or (ii). Now let $\mathcal{R}_i = \mathcal{E} \cup \mathcal{U}_{>i}$ and $P^{(i)} = P|_{\mathcal{R}_i}$, the distribution in which the events of $u_1, \ldots, u_i$ are marginalized, so that $P^{(0)} = P$ and $P^{(l)} = P|_{\mathcal{E}}$. By lemma C.7, each $P^{(i)}$ has a competing events representation, and the clocks that appear in assumption 4.1 for u_i, namely $q_e(T_e \mid \mathcal{H}_t, \mathcal{H}^{u_i}_t, \mathcal{H}^{\mathcal{U}_{>i}}_t)$, are the clocks of $P^{(i-1)}$, a competing events model over $\mathcal{R}_i \cup \{u_i\}$. We show by induction on i that $\tilde{P}|_{\mathcal{R}_i} = \widetilde{P^{(i)}}$. For $i = 0$ this is the definition of $\tilde{P}$. Assume it holds for $i - 1$. Condition (i) or (ii) of assumption 4.1 for u_i is exactly condition (i) or (ii) of the single-type case for $P^{(i-1)}$, with u_i as the unobserved type and $\mathcal{R}_i$ in the role of $\mathcal{E}$. The single-type case therefore gives $\big(\widetilde{P^{(i-1)}}\big)|_{\mathcal{R}_i} = \widetilde{P^{(i)}}$, and

$$\tilde{P}|_{\mathcal{R}_i} = \big(\tilde{P}|_{\mathcal{R}_{i-1}}\big)|_{\mathcal{R}_i} = \big(\widetilde{P^{(i-1)}}\big)|_{\mathcal{R}_i} = \widetilde{P^{(i)}}.$$

All conditional distributions above are taken on histories of positive probability under $\tilde{P}$, which by overlap have positive probability under P and hence under each $P^{(i)}$. For $i = l$ we obtain $\tilde{P}|_{\mathcal{E}} = \widetilde{P|_{\mathcal{E}}}$. Since $f(\mathcal{H}^y)$ and $\mathcal{H}_t$ are functions of the observed trajectory, this gives $\mathbb{E}_{\tilde{P}}[f(\mathcal{H}^y) \mid \mathcal{H}_t] = \mathbb{E}_{\widetilde{P|_{\mathcal{E}}}}[f(\mathcal{H}^y) \mid \mathcal{H}_t]$. □

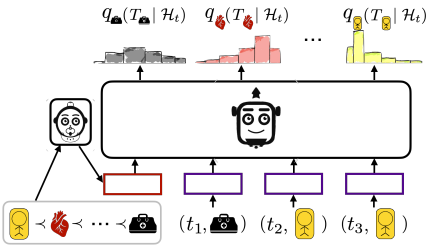


Figure 6: Transformer architecture for CEMs. A small cross-attention module embeds an ordering into prefix tokens (red), which are processed with event descriptions to produce q_e

D Details on CEM Implementation, Simulations and Additional Results

D.1 Details on Implementation

As described in section 4.3, fig. 6 described the architecture that embeds an ordering $\prec$ over the events. Algorithm 2 also gives the pseudo-code for the various stages of the method. In section 4.3 we pointed out that continuous times can be used without implementing integration. This is due to the stochastic estimator of the gradient for the CDF,

$$\nabla_\theta \log 1 - F_e[t_i \mid \mathcal{H}_{t_{i-1}}; \theta] = \nabla_\theta \log \int_{t_i}^{T} q_e(T_{\text{next}} = t \mid \mathcal{H}_{t_{i-1}}; \theta) =$$

$$\frac{1}{\int_{t_i}^{T} q_e(T_{\text{next}} = t \mid \mathcal{H}_{t_{t-1}}; \theta) dt} \int_{t_i}^{T} \nabla_\theta q_e(T_{\text{next}} = t' \mid \mathcal{H}_{t_{i-1}}; \theta) dt' =$$

$$\int_{t_i}^{T} \frac{q_e(T_{\text{next}} = t' \mid \mathcal{H}_{t_{i-1}}; \theta)}{\int_{t_i}^{T} q_e(T_{\text{next}} = t \mid \mathcal{H}_{t_{t-1}}; \theta) dt} \nabla_\theta \log q(T_{\text{next}} = t' \mid \mathcal{H}_{t_{i-1}}; \theta) dt' =$$

$$\mathop{\mathbb{E}}_{t' \sim q_e(\cdot | T_e > t_i; \theta)} \left[\nabla_\theta \log q_e(t' | \mathcal{H}_{t_{i-1}}; \theta) \right]$$

Algorithm 2 CEMs Learning and Sampling

Note: we reintroduce features z_j omitted from our formal treatment.

Sampling from $P(\cdot|\mathcal{H}_t; \boldsymbol{\theta})$

Input: CEM $\{q_e\}_{e \in \mathcal{E}}$, Ordering $\prec$
Sample $t_e \sim q_e(T_e \mid \mathcal{H}_t; \boldsymbol{\theta})\ \forall e \in \mathcal{E}$
Set $e_{\text{next}} \leftarrow \arg\min_{e; \prec} t_e$, $t_{\text{next}} \leftarrow t_{e_{\text{next}}}$
Sample $\mathrm{z}_{\text{next}} \sim q(\mathrm{z} \mid \mathcal{H}_t, t_{\text{next}}, e_{\text{next}})$
Return $\mathcal{H}_t \cup (t_{\text{next}}, e_{\text{next}}, \mathrm{z}_{\text{next}})$

Estimating conditional effects

Input: Outcome $f : \Omega \to \mathbb{R}$, CEM $\{q_e\}_{e \in \mathcal{E}}$, policy $\tilde{q}_a$, history $\mathcal{H}_t$, size $m \in \mathbb{N}$
Replace q_a by $\tilde{q}_a$; set $\prec$ with a last
for $i = 1 \ldots m$ **do**
  $\mathcal{H}_i \leftarrow \text{Sample}(\{q_e(\cdot \mid \mathcal{H}_t)\}, \prec)$
Return $\frac{1}{m} \sum_{i=1}^{m} f(\mathcal{H}_i)$

Training $\{q_e(\cdot; \theta)\}_{e \in \mathcal{E}}$ **Input:** Trajectories $\{\mathcal{H}_i\}_{i=1}^{N}$

for D rounds **do**
  Draw batch $B \subset [N]$ and a random order $\prec$ over $\mathcal{E}$
  Take gradient step on $-\frac{1}{|B|} \sum_{i \in B} \sum_{(t_j, e_j, \mathrm{z}_j) \in \mathcal{H}_i} \left[\mathcal{L}(t_j, e_j \mid \mathcal{H}_{i,t_{j-1}}; \theta) + \log q(\mathrm{z}_j \mid \mathcal{H}_{i,t_{j-1}}, t_j, e_j; \theta) \right]$

D.2 Architecture and Hyperparameters

Details on trained transformers. We trained CEMs on A100 and V100 GPUs, wall clock time comparisons for a MIMIC-IV run is given in the next section. All methods are trained with decoder-only transformers from the x-transformers package with RoPE. For the cancer simulation we tuned learning rate and dropout for all methods over 6 randomly drawn pairs of values in the range $[5e-5, 5e-3]$ for learning rate and $[0, 0.3]$ for dropout. All networks for this setting have 6 layers and 6 attention heads, with a token dimension of 128. For MIMIC-IV, hyperparameter tuning for

baselines was done by (Oufattole et al., 2024), while for CEMs we did not scan hyperparameters and used a learning rate of 0.0003. The token dimension for all methods here is 128 as well. The maximum sequence length is set to 128 for baselines, while for CEMs each event can span 2 tokens since we predict numeric values conditioned on the code they are associated with. Thus we allow a sequence length of 256 for the transformer, but we cap the number of events at 128 so the overall events that all methods observe are the same. In this setting the models for all methods have 2 layers and 4 attention heads. Embedding of tokens (t, e, z) for all methods use a codebook embedding for the categorical values t, e and a Continuous Value Encoder for z. These embeddings are summed to create the token embedding.

For CEMs we embed the ordering $\prec$ as prepended tokens to the sequence. In the cancer simulator, since there are only 4 event types, we embed the 4 tokens and embed their corresponding ordering using another codebook embedder, and sum them. These 4 tokens are prepended to the sequence. For MIMIC-IV we use cross attention with 8 heads and 8 learned query vectors, to embed the ordering over $83,065$ codes into 8 tokens prepended to the sequence. The ordering is drawn uniformly from the set of all permutations for each training batch. We also subsample $K = 1024$ codes at training time for the calculation of the loss as described in the main paper. The real feature values z for events in CEMs are predicted with a linear head over the latent representations, and so are latent time distributions (with $83,065$ heads for MIMIC-IV and 4 heads for the simulated setting).

D.3 Cancer Simulation

Cancer simulator: The tumor growth simulation we use is adapted from Bica et al. (2020); Seedat et al. (2022); Melnychuk et al. (2022) and is based on the work of Geng et al. (2017). Tumor volumes $V(t)$ are simulated as finite differences from the following differential equation,

$$\frac{dV(t)}{dt} = \left(\underbrace{\rho \log\left(\frac{K}{V(t)}\right)}_{\text{Tumor growth}} - \underbrace{\beta_c C(t)}_{\text{Chemotherapy}} - \underbrace{\left(\alpha_r d(t) + \beta_r d(t)^2\right)}_{\text{Radiotherapy}} + \underbrace{e_t}_{\text{Noise}} \right) V(t).$$

Here $C(t)$ is the chemotherapy concentration, $d(t)$ represents the level of radiotherapy. $\rho, K, \beta_c, \alpha_r, \beta_r$ are effect parameters drawn for each patient from a prior distribution described in Geng et al. (2017), and $e_t \sim \mathcal{N}(0, 0.0001)$ is a noise term. To create irregularly sampled observations of the tumor volume, at each time step we draw a value from a Bernoulli distribution to decide whether the trajectory contains the tumor volume at this time step or not. The success probability is a function of the average tumor volume over the most recent 15 volumes (both observed and unobserved). If we denote a missing value by $\emptyset$ and the observation at timestep t by X_t (which equals $\emptyset$ if there is no sample at this timestep and $V(t)$ otherwise), then sampling times are drawn according to the following probabilities:

$$p(V_t \neq \emptyset | \mathcal{H}_t) = \sigma\left(\frac{\bar{V}_{t-15:t}}{V_{\max}} - 1.5\right)$$

The policies we use to decide on treatments draw binary decisions of whether or not to apply chemotherapy and radiotherapy at each timestep. Denoting these decisions by random variables C_t and d_t, they are drawn according to $P(C_t = 1 | \mathcal{H}_t) = \sigma\left(\gamma \cdot (v_{\text{last}} - \beta) + t - t_{\text{last}}\right)$, where v_{last} is the last observed volume before time t and t_{last} is the last time that treatment was applied before t. The same probabilities are applied for d_t.

Baselines: FQE (Le et al., 2019) and EDQ (Wald et al., 2025) both train functions $Q(\mathcal{H}_t)$ that aim to approximate the estimand $\mathbb{E}_{\tilde{P}}[f(\mathcal{H})]$ and are trained via dynamic programming with value-networks. FQE works in discrete time and evaluates a policy $\pi(a_{t+1} \mid \mathcal{H}_t, \mathrm{x}_{t+1})$ where x are vectors of features, and a is the action. It is based on the tower property of conditional expectations $\mathbb{E}_{\tilde{P}}[f(\mathcal{H}) \mid \mathcal{H}_t] = \mathbb{E}_{\mathrm{x}_{t+1} \sim P(X_{t+1} \mid \mathcal{H}_t)}[\mathbb{E}_{a_{t+1} \sim \pi(A_{t+1} \mid \mathcal{H}_t, \mathrm{x}_{t+1})}[f(\mathcal{H}) \mid \tilde{\mathcal{H}}_{t+1} = \mathcal{H}_t \cup (\mathrm{x}_{t+1}, a_{t+1})]]$. The inner expectation on the right hand side is itself a Q-function, hence the algorithm regresses $Q(\mathcal{H}_t)$ onto $Q(\tilde{\mathcal{H}}_{t+1})$, where a_{t+1} is sampled from π, the policy we wish to evaluate, and x_{t+1} is taken from the training dataset as it is a sample from $P(X_{t+1} \mid \mathcal{H}_t)$ which appears in the outer expectation of the right-hand-side. The pseudocode below summarizes the method. FQE is impractical when the horizon is long and events occur irregularly, since to process the data without losing precision

the timeline must be finely discretized, yet fine discretization creates noisy gradients since there is very sparse reward signal, and this keeps the signal from propagating all the way to the beginning of the trajectory. This problem is apparent in our simulation, and would be made much worse in real datasets such as MIMIC-IV where the smallest time gap between events could be close to a second and the largest days or months. Problems with such short horizons motivated N-step methods in RL, and in the context of irregular time and interventions of timing it motivated EDQ.

Algorithm 3 Fitted Q-Evaluation (discrete time)

Input: Trajectories $\{\mathcal{H}_i\}_{i=1}^N$,
Policy $\pi(\mathcal{H}_t, \mathrm{x}_t) \to \Delta^{|A|}$
Initialize $Q(;\theta)$ randomly
Set $T_i \leftarrow \max\{t : (t, \mathrm{z}) \in \mathcal{H}_i\} \quad \forall i \in [N]$
for D rounds **do**
  Draw $\tilde{a}_{t+1} \sim \pi(\cdot|\mathcal{H}_{i,t}, \mathrm{x}_{i,t+1}) \quad \forall i, t \in [T_i]$
  $\tilde{\mathcal{H}}_{i,t+1} \leftarrow \mathcal{H}_{i,t} \cup \{(t+1, \mathrm{x}_{t+1}), (t+1, \tilde{a}_{t+1})\}$
  Set $\hat{y}_{i,t} = y_{i,t} + Q(\tilde{\mathcal{H}}_{i,t+1}) \quad \forall i, t \in [T_i]$
  $\theta \leftarrow \arg\min_\theta \sum_{i \in [N], t \in [T_i]} \left(Q(\mathcal{H}_{i,t}; \theta) - \hat{y}_{i,t}\right)^2$
Return $Q(\cdot;\theta)$

EDQ (Wald et al., 2025) extends the conditional expectation identity we saw above for FQE, so instead of looking one timestep ahead (i.e. updating $Q(\mathcal{H}_t)$ using $Q(\tilde{\mathcal{H}}_{t+1})$ as in the algorithm above), it finds the first time point where the policy we wish evaluate disagrees with the observed trajectory $\mathcal{H}_i$. This alleviates the long-horizon proble since it practically "skips" steps where no update is needed and prevents noisy gradients from interrupting the dynamic programming optimization. However, it is a rather complex method that is difficult to train and scale in practice, and as we see in the empirical evaluation it can underperform w.r.t simpler methods such as CEM. EDQ is formalized with a policy λ_a similar to our definition of a policy in this paper, and λ_a can be equivalently replaced with a distribution q_a that gives the time of the next action of type a unless some other event comes beforehand. We refer to (Wald et al., 2025) for pseudocode, notation and the development of the method.

Table 3: Competing-events NRMSE averaged over simulation start times. Each row corresponds to networks trained on training distribution with a specified policy, while in each column we plug in an evaluation policy to the trained network and estimate its forecasts under data generated with the evaluation policy.

Train / Eval	(10, 0.5)	(10, 0.75)	(6,0.5)	(6, 0.75)
(10, 0.5)	0.232	0.303	0.213	0.347
(10, 0.75)	0.237	0.212	0.184	0.289
(6, 0.5)	0.251	0.236	0.180	0.353
(6, 0.75)	0.259	0.273	0.209	0.288

Table 4: NRMSE when providing models with the history up to the last event that occurs before time t, for different values of t. All methods were trained with data generated from policy $(\gamma, \beta)_{\text{obs}} = (10, 0.75)$ and tested on data generated with policy $(\gamma, \beta)_{\text{int}} = (6, 0.5)$

Method / t	1	4	7	10	13	16	18
FQE	1.113	1.024	0.926	0.807	0.599	0.279	0.201
EDQ	0.330	0.334	0.288	0.241	0.222	0.216	0.206
CEM	0.314	0.259	0.203	0.162	0.147	0.119	0.083
ERM/MC	0.410	0.417	0.373	0.314	0.269	0.255	0.246

Additional Ablations: As an additional ablation on the simulated dataset, we train networks on data generated from one treatment policy (train policies, rows), plug-in an evaluation policy to the trained

model and test on data generated using that evaluation policy. Ideally each column will have nearly identical numbers that are close to the optimal achievable error on the dataset. The results in table 3 show that in practice the trained model we use has an effect on the obtained result, however, as can be seen by comparing to results in table 1 of the main paper, these deviations are small w.r.t to the error obtained by not accounting for the distribution shift between policies. Finally, table 4 gives the average NRMSE of each method when the prediction point t changes. The prediction point for each trajectory is set as the last time where there is an observation on the trajectory prior to the prediction point t. Unsurprisingly, we see that errors grow for all methods as we predict over a longer horizon (i.e. smaller t). FQE's error rises more quickly with the length of the horizon, showing that gradients do not propagate backwards through the trajectory, while EDQ does better but still suffers from bias. While we believe that further finetuning of the update scheme for the gradients (e.g. taking gradient steps at the end and going backward in time as optimization continues) can improve performance, this method requires more fine-tuning per task and is rather inflexible.

D.4 Consistency Across Orderings

The CEM in our experiments is trained with random orderings $\prec$ encoded as prefix tokens (section 4.3), so it should represent the same distribution regardless of the ordering it is given. We test this on the tumor growth simulator in two ways.

Validation loss across orderings. We draw 30 random orderings over the event types and compute the validation loss of the trained model under each of them. Besides the 4 event types, the implementation includes a padding type that never appears in the data, and we draw orderings over it as well. Table 5 reports the mean and standard deviation across orderings of both components of the loss. The low variance indicates that the ordering-conditioned network is internally consistent.

Table 5: Validation loss of the CEM across 30 random orderings (mean $\pm$ std).

Time + event loss (NLL)	Tumor volume prediction (MSE)
0.0891 ± 0.0003	0.0529 ± 0.0043

Effect estimation under different valid orderings. For the intervention in section 6, an ordering is valid if it places the two treatment types last, which gives $3! \cdot 2! = 12$ valid orderings. The static event always precedes all others and the padding type never appears, so the only change among valid orderings that affects tie-breaking on observed data is swapping the two treatments. This determines which treatment wins when chemotherapy and radiotherapy are concurrent, and is therefore the comparison most likely to change the estimate. Table 6 compares the errors of the ordering used in the paper and the swapped ordering, per prediction time as in table 5, with standard deviations from bootstrapping the Monte Carlo estimates. The difference in total NRMSE (0.189 vs. 0.184) is well within one standard deviation.

Table 6: NRMSE of effect estimates per prediction time t under two valid orderings (mean $\pm$ bootstrap std).

Prediction time t	1	4	7	10	13	16	18	Total NRMSE
Ordering from paper	0.314 ± 0.020	0.259 ± 0.018	0.203 ± 0.012	0.162 ± 0.008	0.147 ± 0.007	0.119 ± 0.006	0.083 ± 0.006	0.189 ± 0.009
Swapped treatment order	0.300 ± 0.021	0.261 ± 0.014	0.209 ± 0.010	0.173 ± 0.008	0.147 ± 0.008	0.117 ± 0.006	0.082 ± 0.006	0.184 ± 0.010

The degree of freedom left by the choice of ordering is thus small in practice for this model. For any trained CEM, it can be checked with the conversion procedure in lemma C.7, which can also enforce consistency by training the model to match the converted distributions.

D.5 Finite Sample Evaluation

To study how the accuracy of effect estimates depends on how well the CEM fits the observational distribution, we train CEMs on growing subsets of the tumor growth training set and evaluate them on the intervention of section 6. Table 7 reports the total NRMSE of the interventional estimate (mean $\pm$ bootstrap std) together with the observational validation loss (likelihood plus MSE on the

prediction of x_{t+1}). The estimation error decreases as the training set grows and the observational loss decreases. The effect is most pronounced at small sample sizes; at lower observational losses, other factors, including the variance induced by the policy shift, dominate.

Table 7: Interventional estimation error and observational loss as a function of training set size.

Training set size	500	1000	2000	8000	12000 (full)
Interventional estimand error (total NRMSE)	0.739 ± 0.017	0.474 ± 0.017	0.338 ± 0.020	0.224 ± 0.012	0.189 ± 0.009
Observational loss (validation)	0.277	0.188	0.150	0.146	0.141

D.6 Time-Shifting Baseline

We compare CEMs to a natural baseline for interventions on time in standard autoregressive models. We train a model that predicts $P(E_{\text{next}} \mid \mathcal{H}_t)$ and $P(T_{\text{next}} \mid E_{\text{next}}, \mathcal{H}_t)$, and whenever generation samples a treatment, we replace its predicted time with a time drawn from the target policy $\tilde{q}_a$. This estimator is incorrect because it intervenes on $P(T_{\text{next}} \mid E_{\text{next}})$ rather than on the treatment clock $q_a(T_a)$, which is the failure described in section 2.2: $P(E_{\text{next}} \mid \mathcal{H}_t)$ remains invariant under the intervention.

As in table 1(a), the *in-distribution* column reports the error of models trained and evaluated under the policy $(6, 0.5)$ with no intervention, which measures predictive quality. In the *under shift* column, models are trained under $(10, 0.75)$, and the target policy is applied at inference by time replacement for the autoregressive model and by replacing q_a for the CEM. Table 8 reports total NRMSE with bootstrap standard deviations. The baseline's error increases by 64% under the intervention, while the CEM's is essentially unchanged. The baseline also underperforms ERM/MC (0.327 in table 1(a)), which ignores the policy shift entirely. Since the baseline is also weaker in-distribution (0.250 vs. 0.189), the comparison is best read as the change under the intervention rather than in absolute terms.

Table 8: Total NRMSE of the time-shifting baseline and CEM, in-distribution and under a policy shift (mean $\pm$ bootstrap std).

	In-distribution	Under shift
AR + time replacement	0.250 ± 0.009	0.411 ± 0.013
CEM	0.189 ± 0.009	0.184 ± 0.011

E Additional MIMIC-IV Experiments

Runtime comparisons: A natural concern with CEMs is that introducing per-event-type prediction heads along with ordering networks adds computational overhead relative to a standard autoregressive (AR) model that predicts the next event with a single softmax over $\mathcal{E}$. To quantify this overhead, we compare CEM and AR training runtimes across four model scales on MIMIC-IV. Both models share the same decoder-only transformer backbone and are trained on identical data with the same batch size and sequence length; the only architectural difference is that CEM prepends ordering tokens and augments the backbone with $|\mathcal{E}|$ low-rank prediction heads for the latent times T_e, with $K = 1024$ event types subsampled per step as described in Section 4.3. We measure wall-clock time per training step and peak GPU memory on a single A100 GPU, averaged over training steps after a brief warmup.

Table 9 reports the runtime results. As expected, CEM has higher parameter counts and memory footprint than AR at every scale, since the per-type heads add parameters proportional to $|\mathcal{E}| \cdot d$, where d is the model dimension. However, the relative wall-clock overhead decreases substantially as the backbone grows: from $2.17\times$ at the smallest configuration (4 heads, 2 layers, dim 128) to $1.21\times$ at the largest (16 heads, 16 layers, dim 1024). This trend is intuitive, the per-type heads contribute a fixed cost per step that depends on K and d, while the backbone cost scales with depth and width. As the backbone becomes the dominant compute, the CEM-specific overhead becomes a progressively smaller fraction of the total. We also note that the epochs to model convergence are also similar between CEM and AR even when we subsample a relatively small K. We use the small configuration for the main MIMIC-IV experiments to align with the baselines from MEDS-torch. At the scale of these experiments, the additional cost is modest given the added capability of supporting interventions

Table 9: Runtime and memory comparison between CEM and AR baseline on MIMIC-IV across four model configurations. The configurations are matched in (heads, layers, dim) so that any difference reflects the CEM-specific overhead. Step duration is measured as the mean wall-clock time per training step (forward + backward) on a single A100 GPU; values are reported as mean $\pm$ std over training steps. Peak memory is the maximum GPU memory observed during training. The relative duration column reports the CEM step duration divided by the AR step duration for each configuration.

Model	Configuration			Params	Step duration (s)	Peak GPU memory (GB)	Relative duration
	Heads	Layers	Dim				
AR	4	2	128	43.1M	0.050 ± 0.001	11.5	$2.17\times$
CEM	4	2	128	177.5M	0.108 ± 0.004	16.7	
AR	8	8	512	195.6M	0.095 ± 0.001	15.1	$1.54\times$
CEM	8	8	512	586.2M	0.146 ± 0.004	24.2	
AR	12	12	768	340.8M	0.143 ± 0.002	19.3	$1.34\times$
CEM	12	12	768	902.8M	0.191 ± 0.005	32.1	
AR	16	16	1024	542.7M	0.213 ± 0.001	25.2	$1.21\times$
CEM	16	16	1024	1.28B	0.258 ± 0.007	41.8	

on event timing, and this gap will only get smaller as we increase model sizes to match health-system scale datasets for real-world applications and deployments.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We state results that are demonstrated empirically or by math. The abstract and introduction accurately present the model we develop and experiments we perform. We do not claim results beyond what we show; for example, we do not claim to show reliable causal inference on real-world data.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: While we did not have enough space for a dedicated limitations section, limitations are discussed across the paper. We state identifiability limitations which are crucial and are based on a no-unobserved-confounding and overlap assumption, that does not hold in many real-world scenarios. We also state some added computational burdens and the need for further dedicated work exploring the model on real-world data.

   Guidelines:

- The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: We provide complete proofs in the supplemental material.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We provide most details in the main text and by referring to repositories that we used, though we plan to release the code upon publication and that will enable complete reproducibility.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.

- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: We plan to release the code upon publication, yet it needs to undergo some cleaning and removal of any identifying traces before it is released to reviewers.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).

- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: While some details are left for the appendix, we do state the details of our training procedure. Any detail we may have missed will be completed upon request.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We provide standard deviations either based on bootstrapping (for cancer simulations) or experiments with different random seeds (MIMIC fine-tuning).

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
   - If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We state the use of A100 GPUs, and the limitations regarding memory that require subsampling events. We also provide wall-clock runtimes per-step as measured by a profiler, and state the overall training time for experiments on MIMIC-IV.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.

- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: This paper is prepared in accordance to the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: While the paper states the possible contribution to decision-support tools, which is a positive societal impact, we also use the conclusion section to warn from bias amplification and irresponsible use.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: This study does not release any assets with a high risk for misuse. Data used in our experiments is either synthetic or publicly available under credentialed access controls.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: We cite the MEDS standard and MEDS-torch codebase that our MIMIC-IV experiments build upon. We also reference the works that developed and applied the cancer simulator that we use in experiments.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [N/A]

Justification: While our source code is not released at this time with the submission, it will be properly documented upon release.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.

- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: This work does not involve any crowdsourcing or research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: We use the publicly available MIMIC-IV dataset and therefore assume that appropriate risk disclosure and IRB approvals were completed by the authors of the dataset.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: LLMs have only been used for editing, writing, and formatting purposes.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.