Journal of the Royal Statistical Society Series B: Statistical Methodology, 2025, **87**, 74–97
https://doi.org/10.1093/jrsssb/qkae056
Advance access publication 10 July 2024
Original Article

ROYAL STATISTICAL SOCIETY DATA | EVIDENCE | DECISIONS
Journal of the Statistics Society Series B Statistical Methodology

Graphical criteria for the identification of marginal causal effects in continuous-time survival and event-history analyses

Kjetil Røysland[1], Pål C. Ryalen[1,2], Mari Nygård[3] and Vanessa Didelez[4,5]

[1]Department of Biostatistics, University of Oslo, Oslo, Norway
[2]Department of Mathematics, EPFL, Lausanne, Switzerland
[3]Department of Research, Cancer Registry of Norway, Norwegian Institute of Public Health, Oslo, Norway
[4]Department of Mathematics and Computer Science, University of Bremen, Bremen, Germany
[5]Leibniz Institute for Prevention Research and Epidemiology – BIPS, Achterstr 30, D-28359 Bremen, Germany

Address for correspondence: Vanessa Didelez, Leibniz Institute for Prevention Research and Epidemiology – BIPS, Achterstr 30, D-28359 Bremen, Germany. Email: didelez@leibniz-bips.de

Abstract

We consider continuous-time survival and event-history settings, where our aim is to graphically represent causal structures allowing us to characterize when a causal parameter is *identified* from observational data. This causal parameter is formalized as the effect on an outcome event of a (possibly hypothetical) intervention on the intensity of a treatment process. To establish identifiability, we propose novel graphical rules indicating whether the observed information is sufficient to obtain the desired causal effect by suitable reweighting. This requires a different type of graph than in discrete time. We formally define causal semantics for the corresponding dynamic graphs that represent local independence models for multivariate counting processes. Importantly, our work highlights that causal inference from censored data relies on subtle structural assumptions on the censoring process beyond independent censoring; these can be verified graphically. Put together, our results are the first to establish graphical rules for nonparametric causal identifiability in event processes in this generality for the continuous-time case, not relying on particular parametric survival models. We conclude with a data example on Human papillomavirus (HPV) testing for cervical cancer screening, where the assumptions are illustrated graphically and the desired effect is estimated by reweighted cumulative incidence curves.

Keywords: causal inference, cervical cancer, independent censoring, local independence models, reweighting, survival analysis

1 Introduction

Survival analysis is a fundamental field of biostatistics. The typical aim of many medical or epidemiologcial studies is to investigate, e.g. how to delay the event of death, progression of disease or other untoward occurrences. When the research question is about the behaviour under certain *changes* to the processes, e.g. due to some intervention or manipulation, we consider this as *causal inference*, in accordance with a wide literature (Dawid & Didelez, 2010; Pearl, 2009; Peters et al., 2016; Robins, 1986; Rubin, 1974; Spirtes et al., 2000).

Adopting the framework of counting processes (Aalen et al., 2008; Andersen et al., 1993; Røysland, 2011), we here provide a novel formal and graphical framework for causal reasoning about event-histories in continuous time. The proposed causal notion relies on formalizing the intended interventions as modified intensities of the relevant continuous-time processes. This

Received: February 1, 2022. **Revised:** April 13, 2024. **Accepted:** April 24, 2024

reflects, for example, early versus late treatment initiation, or higher versus lower frequency, say, of radiation therapy (Ryalen, Stensrud, Fosså et al., 2018). However, while evaluating the effect of such an intervention may be the ultimate aim, we here focus on a key requirement for any causal analysis which is to establish whether valid inference about the interventional situation can, at least in principle, be drawn from the available data; latent processes may pose an obstacle, even in randomized studies, especially when they induce unobserved time-dependent confounding (Robins, 1986). Hence, nonparametric identification of the target of inference should be ensured (Manski, 2003; Shpitser & Pearl, 2006). Formulating and checking assumption for identification has benefited hugely from graphical representations, specifically causal directed acyclic graphs (DAGs) (Dawid, 2002; Pearl, 1995; Robins, 2001), where a complete graphical characterization of identifiability based on d-separations is available (Shpitser & Pearl, 2006, 2008). While these have been extended to discrete-time settings with time-dependent treatments (Dawid & Didelez, 2010; Pearl & Robins, 1995), neither causal DAGs nor these criteria can easily be transferred to continuous-time situations modelled with stochastic processes, exhibiting feedback and censoring (Aalen et al., 2012). To remedy this shortcoming, the so-called local independence graphs and the notion of δ-separation have been suggested as alternative representations of (in)dependencies between processes (Didelez, 2006, 2008); so-far, these have lacked an explicit causal semantic despite being used in causal contexts (Mogensen & Hansen, 2020; Røysland, 2012).

The novel and central contribution of the present paper is the concept of *causally valid local independence graphs*, and *a graphical criterion for nonparametric identification of causal parameters*, i.e. aspects of the interventional distribution of the continuous-time event processes. We introduce a general notion of 'eliminable' processes which can be marginalized without destroying the causal structure and which can be checked graphically. While the ensuing conditions for nonparametric identification essentially demand the absence of unobserved (time-dependent) confounding, *censoring* presents an added complication as it further limits, and possibly biases, the information provided by the observable data. As we show, censoring must not only be independent in a probabilistic sense (Andersen, 2005) but also in a causal sense (to be formalized later), i.e. it must allow inference for a situation where censoring is prevented (Hernán & Robins, 2020). Our results on graphical criteria for identification thus encompass issues of time-dependent confounding as well as right-censoring.

Once it has been established that the desired causal parameter is identifiable, we propose to fit a *continuous-time marginal structural model (MSM)*, which can be done nonparametrically by a weighting method (Ryalen et al., 2019). Here, the continuous-time weights are similar to inverse-probability of treatment (IPTW) and censoring weighting (IPCW) known from discrete-time MSMs (Hernán et al., 2000; Robins et al., 2000). In fact, the latter can be considered as an approximation to the continuous-time weights. The weighting approach is based on a change of measure technique known from stochastic calculus and financial modelling (Røysland, 2011).

The outline of our paper is as follows. We begin with some background, focusing on local independence as a dynamic notion of independence and its graphical representation. Then, we propose our central notion of causal validity including an example how it may fail. Proposition 1 establishes the key role of reweighting with a likelihood-ratio process to obtain the interventional distribution. The central result providing graphical rules for nonparametric identifiability is given in Section 5. Proposition 2 then gives general sufficient (graphical) conditions, essentially only assuming that intensities exist, for causal validity in the context of censoring. Section 7 joins all these pieces for the main result on how to fit a marginal structural model using suitable reweighting provided a sufficient set of covariate processes. We conclude by an illustration of our approach with the example of Human papillomavirus (HPV) testing for cervical cancer screening; this is a more advanced analysis, and provides a formal justification for the analysis in Nygård et al. (2014).

1.1 Set-up and notation

We consider a collection of independent individuals, say, patients, observed over time $t \in [0, T]$, where T is end of follow-up, and $t = 0$ denotes baseline. There may be baseline measurements, but events of interest occur over time: the outcome event(s), e.g. survival; a treatment event, e.g. start or switch of treatment; and possibly other events such as the occurrences of side-effects. A special event is censoring. All events are represented by counting processes, where N_t^i indicates how often

an event i has occurred by time t. Where relevant, we specifically denote by N^y and N^x (omitting the subscript t) the outcome and treatment processes, respectively. The process N^c denotes the counting process for censoring; sometimes we refer to C as the jumping time of N^c. We assume throughout that the compensators of the considered counting processes are absolutely continuous, implying that the counting processes jump at different times (with probability 1) and form a multivariate counting process (Andersen et al., 1993). When referring to *the* intensity of a counting process N we are tacitly referring to a version of the predictable intensity of N. We will repeatedly exploit that these intensities, combined with the distribution over the baseline variables, uniquely determine the distribution P over the considered variables and processes (Jacod & Shiryaev, 2003, Theorem III 1.26).

A continuous-time MSM, as introduced formally in Section 7, models the effect of a strategy for generating N^x on some aspect of N^y. Let P, λ denote the observational distribution and intensities, then we use $\tilde{P}$, $\tilde{\lambda}$ for quantities under the interventional regime enforcing that strategy. For example, $\tilde{S}(t) = \tilde{P}(N^y_t = 0)$ might denote the survival probability under a strategy such as 'early treatment initiation' specified by a given intensity $\tilde{\lambda}^x$ for N^x, where $\tilde{\lambda}^x$ is enforced by the intervention. In order to be able to evaluate such effects from observational, i.e. noninterventional, data generated from P, time-varying confounding needs to be addressed. We will formulate probabilistic and graphical criteria characterizing which information can be ignored without introducing bias. In order to address questions relating to unobservable variables or events, we will 'move' between different sets of information: $\mathcal{V}$ denotes the set of all variables and processes deemed 'relevant' to a system, including the censoring process; this includes baseline variables $X \in \mathcal{B}$, as well as counting processes $N \in \mathcal{N}$. Here, 'relevant' refers to the requirement that the resulting model is causally valid as formalized in Section 4.2. Throughout, we are concerned with interventions on processes, not in baseline variables; the latter are included to define subgroups of interest, or because they may be needed to adjust for baseline confounding. Note that an intervention on a process does not affect the distribution of baseline variables as a cause has to come before the effect in time.

It may not be possible to observe the complete set of variables and processes, or we may not be interested in all of these. Thus, $\mathcal{V}_0 = \mathcal{B}_0 \cup \mathcal{N}_0$ denotes the subset of baseline variables and processes that we focus on and can observe; these exclude N^c_t as inference typically aims at an uncensored situation (to be formalized). We will sometimes use the notation $\dot{\cup}$ for the union of pairwise disjoint sets. Where appropriate, we explicitly refer to sets of unobserved variables or processes by $\mathcal{U}$, and to sets of observed processes that are not of primary interest as $\mathcal{L}$. Thus, the analysis is marginal over $\mathcal{U}$ and the causal parameters may further be marginal over $\mathcal{L}$. Note also that often we will not distinguish between baseline variables and processes, as a variable can be understood as a constant process on $t > 0$. We will highlight where it is important to distinguish baseline variables from processes.

Throughout, we use $\mathcal{F}$ (or $\mathcal{F}_t$) for a σ-algebra (or filtration) on the full set of information generated by $\mathcal{V}$. To improve readability, we will prefer the notation $\mathcal{F}$ for the filtration (dropping the subscript t) unless time needs to be explicitly referenced. When a σ-algebra (filtration) is generated by a subset $A \subset \mathcal{V}$ of processes we write $\mathcal{F}^A$ ($\mathcal{F}^A_t$), where we also use $\mathcal{F}^{a,b}$ instead of $\mathcal{F}^{\{a,b\}}$. Here, we associate processes/variables with their indices so that the filtration generated by N^a is denoted by $\mathcal{F}^a$ and the filtration generated by N^a and N^b with $\mathcal{F}^{a,b}$, etc. Finally, $P|_{\mathcal{F}^A}$ denotes the restriction of a probability measure to the reduced set of information $\mathcal{F}^A$.

2 Graphical local independence models

Graphical local independence models (or, local independence graphs) have been suggested and investigated by Schweder (1970), Aalen (1987), Didelez (2007, 2008) and Mogensen and Hansen (2020). They can be seen as stochastic processes' counterpart to directed acyclic graph models or Bayesian networks (Lauritzen, 1996): local independence graphs have directed edges, but allow for cycles to represent dynamic feedback. They graphically encode the probabilistic (in)dependence structure between processes. Unlike Bayesian networks, which represent conditional independencies between variables, the independencies between processes (the nodes) are in terms of local independencies. Informally, this means that, at any time, the presence of one process does not depend on the past of another process given some other information on the past; hence this

is an asymmetric notion of independence (Didelez, 2006). In the graph below, for instance, we represent that the process N^a is locally independent of the process N^b but not vice versa:

$$N^a \longrightarrow N^b \,.$$

Note that there would be a *second* directed edge between the two nodes, from N^b to N^a, if the processes were mutually locally dependent on each other, i.e. if there was feedback.

In this section, we give basic background information on local independence graphs before we formalize the corresponding causal semantics in Section 4.

2.1 Intensities and local independence

A key concept is that of the intensity of a counting process N_t, which throughout we assume to exist for all counting processes. It establishes the dependence between the process' present, or short-term prediction, and 'past information'. Here, we review the concept briefly; a precise mathematical treatment requires tools from martingale theory, see Aalen et al. (2008) or Jacod and Shiryaev (2003).

Let $\mathcal{F}$ be a filtration generated by a set of variables and processes, e.g. those in a set $\mathcal{V}$ including N, then the $\mathcal{F}$-intensity of N satisfies

$$E(dN_t|\mathcal{F}_{t-}) = \lambda_t\, dt. \tag{1}$$

Crucially, the intensity depends on what past information we include. For instance, the $\mathcal{F}$-intensity and the $\mathcal{G}$-intensity, for a reduced information $\mathcal{G}_t \subset \mathcal{F}_t$, of the same process N are not necessarily identical; the latter can be obtained through the Innovation Theorem (Andersen et al., 1993, II.4.2).

However, when the intensity for a reduced set of past information does remain the same, we speak of a local independence. Specifically, consider $N^a \in \mathcal{N}$, a set $\mathcal{C}$ of baseline variables and processes excluding a, and a single b indexing a process or a baseline variable. Then we say that N^a is *locally independent* of N^b given $\{a\} \cup \mathcal{C}$ if the $\mathcal{F}^{\{a,b\}\cup\mathcal{C}}$-intensity of N^a is indistinguishable from its $\mathcal{F}^{\{a\}\cup\mathcal{C}}$-intensity. Thus, local independence formalizes the intuitive notion that the short-term prediction of N^a is unchanged when removing information on the past of N^b as long as past information generated by variables and processes in $\mathcal{C}$ and its own past, is given. To make this last point explicit, we always include the process N^a itself in the conditioning set (note, Didelez, 2008 uses the same concept but slightly different notation). Alternatively, we can define local independence using martingales: N^a is locally independent of N^b, given $\{a\} \cup \mathcal{C}$, if $N^a_t - \int_0^t \bar{\lambda}_s ds$ is a local martingale with respect to $\mathcal{F}^{\mathcal{V}}$ when $\bar{\lambda}$ is the $\mathcal{F}^{\{a\}\cup\mathcal{C}}$-intensity of N^a. We write local independence as $N^b \nrightarrow N^a \mid N^{(\{a\}\cup\mathcal{C})}$ or $b \nrightarrow a \mid (\{a\} \cup \mathcal{C})$ for short.

As a convention, we always include a process' own history in the filtrations for its different intensities. A detailed treatment of local independence can be found in Aalen (1987) and Didelez (2008) with further generalizations by Mogensen and Hansen (2020). Some important properties of local independence are (i) that it is asymmetric and (ii) it is relative to the set of given information. If N^a is locally independent of N^b given $\{a\} \cup \mathcal{C}$ then this does not necessarily imply the converse, nor does it imply local independence for a subset $\mathcal{C}_0 \subset \mathcal{C}$.

2.2 Local independence graphs and models

We now turn to the graphical representation of local independence structures. For formal details on local independence graphs and models, see Didelez (2006, 2008).

A graph $G = (\mathcal{V}, \mathcal{E})$ is given by a set of vertices (or nodes) $\mathcal{V}$ and directed edges $\mathcal{E}$; the nodes represent variables or processes; there can be up to two edges between nodes representing dynamic relations. The induced subgraph G_A, $A \subset \mathcal{V}$, has nodes $\mathcal{V} \cap A$ and edges $\mathcal{E} \cap (A \times A)$; a subgraph G'_A on A is given if the edges are a subset of those of the induced subgraph. Any node $b \in \mathcal{V}\backslash\{a\}$ with an edge $b \longrightarrow a$ is called a parent of a, while a is a child of b; graphical ancestors or descendants are defined analogously in terms of sequences of directed edges. In order to represent local independence structures, a graph should satisfy the following properties:

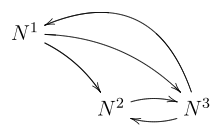
- the node set, $\mathcal{V} = \mathcal{B} \cup \mathcal{N}$, consists of two types, representing either baseline variables or processes;
- all edges are directed;
- between two nodes in $\mathcal{N}$ there may be up to two edges, one in each direction;
- between two nodes in $\mathcal{B}$ there can only be up to one edge;
- there are no edges pointing from a node in $\mathcal{N}$ to a node in $\mathcal{B}$;
- on the subset $\mathcal{B}$ of baseline variables the graph is a DAG.

We call a graph with the above properties *local independence graph*.

A *graphical local independence model*, $(\mathcal{P}, \mathcal{F}, G)$, combines the above graph with a class $\mathcal{P}$ of distributions by demanding that when there is no edge pointing from a given node to a given process then the corresponding local independence must hold for every $P \in \mathcal{P}$. Among the baseline variables (i.e. for their joint marginal distribution), we require the usual directed Markov properties of conditional independence graphs to hold (Lauritzen, 1996). Formally, let G be a local independence graph satisfying the above properties. The corresponding graphical local independence model is a class of joint distributions $\mathcal{P} = \mathcal{P}(G)$ for all possible outcomes or trajectories of the nodes in $\mathcal{V}$ such that, under any $P \in \mathcal{P}(G)$

- the $\mathcal{F}$-intensity (1) is well defined for each counting process in $\mathcal{N}$;
- when $(b \to a) \notin \mathcal{E}$, and $a \in \mathcal{N}$, then N^a is locally independent of N^b given $N^{\mathcal{V}\setminus\{b\}}$;
- $P|_{\mathcal{F}^{\mathcal{B}}}$ satisfies the conditional independencies given by the directed Markov properties of the induced subgraph $G_{\mathcal{B}}$.

Under regularity conditions, the above definition implies that every counting process N^a is locally independent of its nonparents, conditionally on its closure, defined as $\mathrm{cl}(a) = \{a\} \cup \mathrm{pa}(a)$ (Didelez, 2008). This means that the $\mathcal{F}$-intensity of N^a is indistinguishable from its $\mathcal{F}^{\mathrm{cl}(a)}$-intensity. In example (2), we find that N^1 is locally independent of N^2 given $N^{1,3}$, and that there are no other local independencies implied by the graph:



(2)

A further example, combining baseline variables and processes, is given in online supplementary material, Supplement 1.

2.3 δ-separation

In order to use graphical local independence models for causal reasoning, and especially to assess identification of causal parameters, we need to be able to read off independencies retained, or dependencies introduced, when marginalizing over possibly unobservable/latent variables or processes, e.g. to check if these unobservables could induce confounding bias. For instance in the above example graph (2), if N^3 were unobservable it could induce a (marginal) local dependence of N^1 on N^2. This type of property can be read off from a local independence graph by means of δ-separation (Didelez, 2006, 2008), in analogy to d-separation for DAGs. Due to the asymmetric nature of local independence, δ-separation must also be asymmetric and is therefore different from d-separation.

Before formally defining δ-separation, we require the notions of 'blocked trail' and 'allowed trail'. A trail is a subgraph of G formed by unique vertices $\{v_0, \ldots, v_m\}$ and edges $\{e_1, \ldots, e_m\}$ such that either $e_j = v_{j-1} \to v_j$ or $e_j = v_{j-1} \leftarrow v_j \in \mathcal{E}$ for every $j = 1, \ldots, m$. The trail is said to start in v_0 and end in v_m. As there can be multiple edges between nodes, there can be different trails on the same set of nodes. A trail is said to be *blocked* by a set of vertices $C \subset \mathcal{V}$ if either (i) C contains a vertex v_j, $j \in \{2, \ldots, m-1\}$, on the trail such that $e_j = v_j \to v_{j+1}$ or $e_{j-1} = v_{j-1} \leftarrow v_j$ on the trail (i.e. v_j is a noncollider), or (ii) the trail contains the edges $v_{j-1} \to v_j \leftarrow v_{j+1}$ such that C contains

neither v_j nor any of its descendants. Otherwise, the trail is said to be *open* relative to C. An *allowed* trail from a node v_0 to a node v_m in $\mathcal{N}$ is a trail ending with a directed edge into the node v_m, i.e. $e_m = v_{m-1} \to v_m$.

In DAGs, for disjoint sets A, B, C, we say A and B are d-separated by C, if every path between A and B is blocked by C; this separation is symmetric in A and B. For distributions that satisfy the Markov properties of a DAG, every d-separation entails the corresponding conditional independence, i.e. $X_A \perp\!\!\!\perp X_B \mid X_C$ (Lauritzen, 1996; Pearl, 2009, Theorem 1.2.4). In a graphical local independence model, this is still the case for baseline variables $A, B, C \subset \mathcal{B}$, but in addition we use δ-separation to read off local independencies as follows Didelez (2006, 2008).

Definition 1 Let $a \in \mathcal{N}$ and $B, C \subset \mathcal{V}$ be disjoint subsets of vertices in a local independence graph G. Then B is δ-separated from $\{a\}$ by $\{a\} \cup C$ if every allowed trail from any $b \in B$ to a is blocked by C. For a subset $A \subset \mathcal{N}$, B is δ-separated from A by $A \cup C$ if every allowed trail from any $b \in B$ to any $a \in A$ is blocked by $(A \cup C) \backslash \{a\}$. We then write $B \nrightarrow_G A \mid A \cup C$.

The role of δ-separation is in guaranteeing (under regularity conditions) a corresponding local independence in the model (Didelez, 2008, Theorem 1 and 3.4): whenever $B \nrightarrow_G A \mid A \cup C$ then the sub-process A is locally independent of the processes (or variables) in B, given $A \cup C$. As $A \cup B \cup C$ does not need to equal $\mathcal{V}$, δ-separation allows us to infer marginal local independencies in subsets of the system. Note that the 'blocking of allowed trails' condition has an equivalent 'moral graph' condition which allows to check δ-separation on an undirected graph (Lauritzen, 1996); this and additional information on δ-separation can be found in Didelez (2006).

Remark 1 The definition of local independence graph and δ-separation takes as implicit that every process always depends on its own past so that we do not make use of any self-loops and always condition on a process' own past. Mogensen and Hansen (2020) make such a distinction; moreover the authors generalize their treatment to dependencies due to latent processes shown graphically as bi-directed edges and self-loops. The corresponding notion of separation is called μ-separation. See online supplementary material, Supplement 1 for further examples.

Remark 2 The above notion of local independence graph ensures that whenever there is a δ-separation in the graph then there is a local independence in the model $\mathcal{P}(G)$. If the converse holds, i.e. *all* local independencies that occur in every distribution under a probabilistic model $\mathcal{P}$ can be read off from the graph G via δ-separation, then we say that the model is *faithful* to the graph (Meek, 1995), and the graph corresponding to $\mathcal{P}$ is then unique. Faithfulness is especially relevant in the context of causal discovery (Mogensen et al., 2018). For our following results, here, we do not require faithfulness; nevertheless, in slight abuse of terminology, we will simply say 'the local independence graph' when we mean the whole local independence model.

3 Causal validity of local independence models

Graphical local independence models describe the probabilistic dynamic dependence structure of a multivariate counting process (allowing baseline variables). We now combine this with causal semantics. To this end, we need to be explicit about assumptions that link the probabilistic structure with hypothetical interventions under which the data generating process is modified, reflecting, e.g. the situation of earlier treatment initiation. We proceed in analogy to the case of random variables linked by a causal DAG model with a factorization of the distribution and corresponding causal Markov property. Such a causal DAG reflects the causal structure by demanding that the joint distribution obeys the 'manipulation theorem' or 'truncation formula' which is a modification of the factorized distribution (Didelez, 2018; Pearl, 2009; Spirtes et al., 2000). In Section

4.2, we propose an analogous notion for the causal interpretation of local independence graphs in terms of a truncation of the corresponding factorization which we recall first.

3.1 Local characteristics

Let $\mathcal{P}(G)$ be a local independence model on a set of counting processes and baseline variables $\mathcal{V} = \mathcal{N} \cup \mathcal{B}$, with $n_0 = |\mathcal{N}|$ and $n = |\mathcal{V}|$. The joint distribution of $\mathcal{N}$ and $\mathcal{B}$ can be uniquely and explicitly characterized on the so-called canonical space of a marked point process (Jacobsen, 2006; Last & Brandt, 1995). This is the domain space of $\mathcal{B}$ multiplied by the space of all multivariate counting process realizations, so it is not restrictive to take this as the underlying probability space. For instance, if $\mathcal{N}$ has at most m jumps (uniformly bounded) the distribution function factorizes as follows:

$$P(\mathcal{N} \in \mathrm{d}\phi, \mathcal{B} \in \mathrm{d}x) = \prod_{i=1}^{n} Z^i(\mu^i(\omega))\mathrm{d}\phi\nu(\mathrm{d}x), \tag{3}$$

where $\omega = (\phi, x)$, ϕ takes values in a subset of $\mathbb{R}^{mn_0}$, $\mathrm{d}\phi$ is the Lebesgue product measure, and $\mu^1, \ldots, \mu^n$ are the *local characteristics*. The assumption that $\mathcal{N}$ has at most m jumps can then be relaxed by taking limits. The factors in (3) corresponding to counting processes $i \in \mathcal{N}$ can be computed from the $\mathcal{F}$-intensities λ^i and the previous jumps. We have that

$$Z^i(\lambda^i) := \prod_{s_i \leq T} \lambda^i_{s_i} \exp\Big(-\textstyle\int_0^T \lambda^i_s \mathrm{d}s\Big), \tag{4}$$

where s_i denotes the jump times of the counting process N^i.

The graphical structure is reflected in the fact that, as explained earlier, the above $\mathcal{F}$-intensities λ^i are indistinguishable from the $\mathcal{F}^{\mathrm{cl}(i)}$-intensities, i.e. those generated by the past on the graphical parent-nodes of a process and its own past. Additionally, when $i \in \mathcal{B}$, the local characteristics are not functions of time and are given by the conditional probabilities $P(X^i \mid X^{\mathrm{pa}(i)})$ as in the factorized density of a Bayesian network (Lauritzen, 1996).

A simple example of a local independence graph is

$$G : X \longrightarrow N^1 \longrightarrow N^2 \tag{5}$$

where X is a baseline variable and N^1 and N^2 are counting processes. In this case we have that $\mathrm{cl}(1) = \{X, N^1\}$ and $\mathrm{cl}(2) = \{N^1, N^2\}$ (recall the notation from Section 3.2). Thus, the local independencies $X \nrightarrow N^2 | N^{1,2}$ and $N^2 \nrightarrow N^1 | (N^1, X)$ hold for every distribution in the local independence model $\mathcal{P}(G)$. Letting λ^1 and λ^2 denote the $\mathcal{F}^{X,N^1,N^2}$-predictable intensities of N^1 and N^2 with respect to some $P \in \mathcal{P}(G)$, we thus have that λ^1 defines an $\mathcal{F}^{\mathrm{cl}(1)}$-intensity of N^1 and λ^2 defines an $\mathcal{F}^{\mathrm{cl}(2)}$-intensity of N^2 with respect to P. The density (3) is given by the product

$$Z^1(\lambda^1) \cdot Z^2(\lambda^2) \cdot Z^X(\mu^X),$$

where $Z^i(\lambda^i)$ is as in (4) for $i \in \{1, 2\}$, and $Z^X(\mu^X)$ is a density of X with respect to the dominating measure ν.

3.2 Causal validity

In this section, we formalize the notion of causal validity for graphical local independence models. Similar to most of the causal frameworks for random variables and causal DAGs, our definition reflects that some aspects of the system are considered invariant (or stable, or modular) under certain interventions on other parts of the system (Dawid & Didelez, 2010; Pearl, 2009; Peters et al., 2016; Spirtes et al., 2000). More specifically, we consider a hypothetical intervention on process N^i replacing its $\mathcal{F}^{\mathrm{cl}(i)}$-intensity λ^i by a different intensity $\tilde{\lambda}^i$ which is typically assumed to be $\mathcal{F}^{\mathcal{V}_0}$-predictable, e.g. generated by a subset $\mathcal{V}_0$, with the special case where it is predictable just

with respect to its own $\mathcal{F}^i$-history. The latter mimics the case of randomization or exogeneity; for example, in the case where N^i counts the times an individual takes a medication or receives radiotherapy, a possible intervention could be to increase or decrease the frequency regardless of the individual's history. Letting the interventional intensity be $\mathcal{F}^{\mathcal{V}_0}$-predictable allows for dynamic interventions, such as a treatment being (dynamically) intensified after the occurrence of a specific event such as a side-effect.

While the original model $\mathcal{P}$ describes the system's natural behaviour without intervention, we denote with $\tilde{\mathcal{P}}$ the model for the system under such an intervention. The latter may be obtained from the former by simple substitution of the local characteristic of node i in (3) in analogy to the 'truncation formula' (Pearl, 2009; Spirtes et al., 2000) as defined next.

Definition 2 Let $\mathcal{P}(G)$ be a graphical local independence model. Consider an intervention on node $i \in \mathcal{V}$ (or on set of nodes $A \subset \mathcal{V}$). We define the corresponding *intervention model* $\tilde{\mathcal{P}}(G)$ by replacing (all) μ^i by $\tilde{\mu}^i$ ($i \in A$) while the local characteristics of the remaining nodes remain the same in $\mathcal{P}(G)$ and $\tilde{\mathcal{P}}(G)$. Formally, if the joint density of a specific $P \in \mathcal{P}(G)$ is given by (3), then the corresponding $\tilde{P}$ is obtained as

$$\prod_{j \in V \setminus A} Z^j(\mu^j, t) \prod_{i \in A} Z^i(\tilde{\mu}^i, t). \tag{6}$$

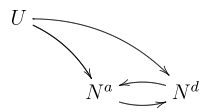
Whenever the above construction of $\tilde{P}$ from P is judged appropriate in a given real-world context, we say that the model is *causally valid with respect to the intervention on node* $i \in \mathcal{V}$ (or $A \subset \mathcal{V}$). Ideally, this would be verified by actually carrying out the desired intervention. In the absence of such experimental validation, subject matter considerations must be used to justify this way of linking the interventional regime $\tilde{P}$ to the observational P in a given real-world context; causal validity will typically only be plausible if the system with its elements $\mathcal{V}$ is sufficiently rich, e.g. in terms of specifying the relevant underlying mechanisms, as illustrated in Section 4.3.

As the above definition is based on interventions that replace intensities, it can be regarded as a 'weak' notion of causality in contrast to 'strong' notions that are based on replacing equations in structural systems, e.g. in causally interpreted stochastic differential equations (Hansen & Sokol, 2014; Mogensen et al., 2018).

A causally valid local independence model without processes, only baseline variables, reduces to a (locally) causal DAG (Pearl, 2009). Further, in the case of variables, we note that different choices for (3) can be probabilistically equivalent but imply different causal relations: while both $P(X^1, X^2) = P(X^1|X^2)P(X^2)$ and $P(X^1, X^2) = P(X^2|X^1)P(X^1)$, only one of the two (if any) factorizations, corresponding to $X_1 \leftarrow X_2$ or $X_1 \rightarrow X_2$, can be causally valid, i.e. either an intervention on X^2 affects X^1 or vice versa. For processes, however, the ordering is explicit in 'time' so that the main issue is whether the past on the included processes and variables in $\mathcal{V}$ contains sufficient information to warrant causal validity.

3.3 Example how causal validity can fail

We consider a simplified clinical situation where a medical condition U occurs with odds γ. This condition can lead to organ failure at some later point in time when the counting process N^d jumps. The patient may receive a treatment N^a to prevent organ failure if U has occurred, as long as he has not experienced organ failure yet. More formally we consider probability densities P such that: $P(U=1)/P(U=0) = \gamma$, and $\lambda_s^a := Y_s \alpha U$ and $\lambda_s^d := Y_s U$ define P-intensities for N^a and N^d with respect to $\mathcal{F}_t^{\{a,d,U\}}$, where $Y_s := I(N_{s-}^a = N_{s-}^d = 0)$. The example is compatible with the following local independence graph:



We furthermore assume that we are able to intervene to prevent the use of this treatment, which means that we hypothetically force $\alpha = 0$. Moreover we assume that our model on all three nodes is causally valid with respect to this intervention, i.e. the odds of U and intensity of N^d remain the same in this hypothetical scenario where the frequencies of events are governed by $\tilde{P}$.

Now assume that the role of U has been overlooked by the analyst, who wrongly thinks that the submodel induced by N^a and N^d, i.e. marginalized over U, is causally valid with respect to an intervention eliminating treatment. To see that this is not the case, we have to show that, when ignoring U, the intensity of N^d is not the same with as without this intervention. To see this, we first apply Bayes formula to obtain

$$P(U=1|Y_s=0) = \frac{\gamma}{\gamma + \dfrac{P(Y_s=0|U=0)}{P(Y_s=0|U=1)}} = \frac{\gamma}{\gamma + e^{s(\alpha+1)}}.$$

A similar argument shows that for the intervened system $\tilde{P}(U=1|Y_s=0) = \gamma/(\gamma+e^s)$. With this we derive the intensity of N^d in the intervened system; formally, let $\tilde{\lambda}^d$ be a version of the $\mathcal{F}_t^{\{a,d\}}$-intensity for N^d w.r.t. $\tilde{P}$. The Innovation Theorem (Andersen et al., 1993, II.4.2) gives that

$$\tilde{\lambda}_s^d = E_{\tilde{P}}[\lambda_s^d|\mathcal{F}_{s-}^{a,d}] = E_{\tilde{P}}[Y_s U|\mathcal{F}_{s-}^{a,d}] = Y_s \tilde{P}(U=1|Y_s=0) = \frac{Y_s\gamma}{\gamma+e^s}, \quad \tilde{P} \text{ a.s.} \tag{7}$$

Without the intervention and still marginally over U, using the same argument as in (7), we find that the corresponding intensity is

$$\bar{\lambda}_s^d = \frac{Y_s\gamma}{\gamma+e^{s(\alpha+1)}}, \quad P \text{ a.s.} \tag{8}$$

Hence, we see that (8) is clearly in conflict with (7), and the induced submodel is not causally valid unless $\alpha = 0$ in the observational scenario, i.e. under P.

The example can be seen as a simple dynamic illustration of confounding: Ignoring the role of U would lead us to under-estimate the risk of organ failure under a no-treatment intervention because, observationally, only low-risk patients tend to remain untreated. The example demonstrates that causal validity is typically only plausible when the multivariate system of processes and variables considered contains sufficient information on 'common causes' even if they are latent, such as the variable U above. In the context of causal DAGs, the assumption that there are no omitted variables is known as 'causal sufficiency' (Hernán & Robins, 2020; Spirtes et al., 2000). As we will see in Section 5, certain variables or processes can, however, be ignored without destroying causal validity. In online supplementary material, Supplement 6 we provide a more academic example where both faithfulness and causal validity are violated.

3.4 Reweighting: likelihood-ratios and positivity

Recall that we want to infer from an observational setting properties under a hypothetical interventional regime. As shown in Proposition 1, reweighting will play a particular role in this endeavour. The hypothetical regime described by $\tilde{\mu}^t$ can in principle be arbitrary and should be chosen to suit the considered, practically relevant, intervention. However, if we want to learn about the hypothetical regime from data obtained under an observational regime, the two cannot be 'too different' from each other. To formalize this we demand absolute continuity $\tilde{P} \ll P$, i.e for every event $H \in \mathcal{F}_T$ with $\tilde{P}(H) > 0$ we also have that $P(H) > 0$. The following proposition shows that this is closely linked to the existence of a likelihood-ratio process which reweights the observational distribution P into the interventional distribution $\tilde{P}$.

Proposition 1 Let $\mathcal{P}(G)$ be a local independence model. Consider a hypothetical intervention on component N^* with $\mathcal{F}^{\mathcal{V}}$-intensity λ^*, and that this intervention

imposes the new intensity $\tilde{\lambda}^* := \rho \cdot \lambda^*$ where ρ is a nonnegative and predictable process. Let

$$W_t := \prod_{s \le t} \rho_s^{\Delta N_s^*} \exp\Big\{ - \int_0^t (\rho_s - 1)\lambda_s^* \mathrm{d}s \Big\}. \tag{9}$$

The following statements are equivalent:

(a) $\mathcal{P}(G)$ is causally valid with respect to an intervention on N^*, and for any $P \in \mathcal{P}$, with corresponding interventional $\tilde{P} \in \tilde{\mathcal{P}}$, we have $\tilde{P} \ll P$ on $\mathcal{F}_T$.
(b) We have that

$$E_{\tilde{P}}(H) = E_P(W_t H) \tag{10}$$

for every $\mathcal{F}_t$-measurable variable H and $t \le T$.

The proposition shows that weighting the events before t according to W_t provides the probabilities in the hypothetical situation under the envisaged intervention. The causal validity of the system ensures the simple structure of W in (9): it only depends on the local characteristic of the intervened node itself.

For a causally valid system, $\tilde{P} \ll P$ if and only if the process W is uniformly integrable on $[0, T]$. There exist several different conditions that imply uniform integrability, see Jacod and Shiryaev (2003, Theorem IV 4.6) and Røysland (2011) or Kallsen and Shiryaev (2002) for more general results.

The conditions that are most relevant for us, translate into upper boundaries on the predictable processes $\rho_t = \tilde{\lambda}_t^*/\lambda_t^*$ and $|\tilde{\lambda}_t^* - \lambda_t^*|$. Such upper boundaries can also be seen as generalization of the positivity condition that is usually assumed (Hernán & Robins, 2020, 3.3). Moreover, the weights (9) can further be regarded as continuous-time version of the *stabilized IPTWs* for discrete-time MSMs (Hernán et al., 2000; Robins et al., 2000). A weaker positivity condition in a similar context, but still only for a discrete-time setting, has been considered by Kennedy (2019).

As a consequence (see also Section 7), when W is known or identified from the observable data, valid statistical analyses of the hypothetical scenario may use weighted averages or weighted regression analyses. It may then be desirable to impose further restrictions on how $\tilde{\lambda}^i$ and λ^i are allowed to differ so that the weights are well-behaved for stable statistical inference (Røysland, 2011, 2012).

4 Causal validity under marginalization

We now turn to the question of when causal quantities are identified even though certain processes are unmeasured, i.e. when the data only carry information on a marginalized system. Identifiability is at the core of many statistical problems, especially missing data, latent variables and causal inference problems (Manski, 2003; Shpitser & Pearl, 2006). In brief, a parameter ξ, being a function of a distribution P in a model $\mathcal{P}$, is said to be identifiable from incomplete information $\mathcal{G}$ if it can uniquely be determined from $P|_{\mathcal{G}}$ for every P in $\mathcal{P}$ (see online supplementary material, Supplement 2 for a formal treatment of identifiability).

A causal parameter ξ is now a function of the interventional distribution $\tilde{P} \in \tilde{\mathcal{P}}$ resulting from replacing the local characteristics of a given node, cf. Definition 2. For example, we may be interested in some aspect of the survival curve under a specific intervention on some treatment process. A causal parameter always induces another parameter $\tilde{\xi}$ in the original model obtained by first mapping P into $\tilde{P}$ using the above likelihood-ratio (9), and then applying ξ, i.e. $\tilde{\xi}(P) = \xi(\tilde{P})$. Thus, a causal parameter is identifiable from incomplete information $\mathcal{G}$ if $\tilde{\xi}$ is identifiable from $\mathcal{G}$.

4.1 Eliminability

The following definition of eliminable processes characterizes graphically when certain subprocesses (or baseline variables) can safely be ignored without destroying certain aspects of the causal

local independence structure. It combines the notions of 'sequential randomization' and 'sequential irrelevance' of Dawid and Didelez (2010). Eliminability is then used to establish our key results on identifiability.

We require some notation, first. Here, as before, we are interested in intervening on a process N^* and consider its effect on a set of outcome *processes* $\mathcal{N}_0$, both exclude baseline variables. Moreover, an intervention on N^* does not affect baseline variables as the intervention takes place after baseline. Hence, the marginal distribution on the baseline variables $\mathcal{B}$ is the same as the corresponding marginal under an intervention, i.e. $P|_{\mathcal{F}^{\mathcal{B}}} = \tilde{P}|_{\mathcal{F}^{\mathcal{B}}}$. However, we allow for baseline variables as well as processes in the set over which we want to marginalize.

Definition 3 Let G be a local independence graph with nodes $\mathcal{V} = \mathcal{V}_0 \dot{\cup} \mathcal{U}$ for $\mathcal{V}_0 = \mathcal{N}_0 \cup \mathcal{B}_0$ with $\mathcal{N}_0 \subset \mathcal{N}$ and $\mathcal{B}_0 \subset \mathcal{B}$; let $N^* \in \mathcal{N}_0$, and $\mathcal{N}_0^{\backslash *} = \mathcal{N}_0 \backslash \{N^*\}$. Then we say that, in G, the set $\mathcal{U}$ is *eliminable* with respect to $(N^*, \mathcal{V}_0^{\backslash *})$ if it can be partitioned into a sequence of sets $\mathcal{U}_1, \ldots, \mathcal{U}_K$ such that, for each $k = 1, \ldots, K$, either

$$\mathcal{U}_k \nrightarrow_G (\mathcal{N}_0^{\backslash *}, \bar{\mathcal{U}}^{k+1}) \,|\, (\mathcal{V}_0, \bar{\mathcal{U}}^{k+1}) \tag{11}$$

or

$$\mathcal{U}_k \nrightarrow_G N^* \,|\, (\mathcal{V}_0, \bar{\mathcal{U}}^{k+1}), \tag{12}$$

holds. Here $\bar{\mathcal{U}}^{k+1} = (\mathcal{U}_{k+1}, \ldots, \mathcal{U}_K)$ and $\bar{\mathcal{U}}^{K+1} = \emptyset$.

Note, when $\bar{\mathcal{U}}^{k+1}$ contains baseline variables, then (11) has to be understood as d-separation between $(\mathcal{U}_k \cap \mathcal{B})$ and $(\bar{\mathcal{U}}^{k+1} \cap \mathcal{B})$ given $(\mathcal{V}_0 \cap \mathcal{B})$ in addition to the δ-separation $\mathcal{U}_k \nrightarrow_G (\mathcal{N}_0^{\backslash *}, \bar{\mathcal{U}}^{k+1} \cap \mathcal{N}) \,|\, (\mathcal{V}_0, \bar{\mathcal{U}}^{k+1})$.

The theorem below shows that causal validity regarding an intervention on N^* is retained when ignoring (i.e. marginalizing over) eliminable sets $\mathcal{U}$. This is related to and a considerable generalization of the 'noninformative treatment assignment' proposed by Arjas and Parner (2004) in the context of marked point processes; the property of eliminability also has some similarity to some principles of the selecting covariates to adjust for confounding in DAGs (VanderWeele & Shpitser, 2011; Witte & Didelez, 2019). An immediate implication is that in case of eliminability, the likelihood-ratio in the subsystem and hence any corresponding causal parameter is identified from $\mathcal{F}^{\mathcal{V}_0}$ (see Section 5.2). When processes are unobservable it can be helpful to reassure ourselves that they are eliminable to ensure identifiability from observables.

Theorem 1 Consider a local independence model $\mathcal{P}(G)$. Let the nodes be partitioned as in Definition 3. Assume causal validity with respect to an intervention on the process $N^* \in \mathcal{N}_0$, replacing its $\mathcal{F}^{\mathcal{V}}$-intensity λ^* by a $\mathcal{F}^{\mathcal{V}_0}$-intensity $\tilde{\lambda}^*$.

If $\mathcal{U}$ is eliminable with respect to $(N^*, \mathcal{V}_0^{\backslash *})$ in G, then the model restricted to $\mathcal{F}^{\mathcal{V}_0}$ (i.e. marginally over $\mathcal{U}$) is also causally valid with respect to the same intervention.

Conditions (11) and (12) are sufficient for processes to be ignored without destroying causal validity. Consider for instance the local independence graph

$$U^1 \to N^* \leftrightarrows N^y \leftarrow U^2,$$

assuming causal validity with respect to an intervention on N^*. First, by δ-separation (11) we see that N^y is locally independent of U^1 given (N^*, U^2, N^y). Second, with (12), we find that N^* is locally independent of U^2 given N^y, N^*. Hence, the submodel on (N^*, N^y) is causally valid with respect to an intervention on N^* as long as this intervention does not depend on (U^1, U^2); the model essentially asserts that there is no confounding of (N^*, N^y) regardless of whether (U^1, U^2) are observed.

Further basic examples where U can be ignored, are described by the local independence graphs

$$N^* \leftarrow U \leftarrow N^y \qquad N^* \rightarrow U \rightarrow N^y.$$

In the first case condition (11) holds, in the second condition (12) of Theorem 1. Here, U could also be a sequence of such processes on directed paths. More examples for local independence models with different structures of eliminable processes can be found in online supplementary material, Supplement 1.

Remark 3 Our result on eliminability is related to the marginalization considered by Mogensen and Hansen (2020). The authors propose an extended class of local independence graphs, and corresponding μ-separation, which is closed under marginalization. These more general graphs include bi-directed edges as a possible result of latent processes not shown as nodes in the graph. In our case, if we consider $\mathcal{U}$ as latent processes and if they satisfy the conditions of eliminability, then they do not induce any bi-directed edges with endpoints between N^* and $\mathcal{V}_0^{\setminus *}$ in these more general graphs. Moreover, their results can be used to obtain the 'latent projection' graph representing the local independence structure after marginalizing over $\mathcal{U}$ (Mogensen & Hansen, 2020, Definition 2.23 and Theorem 2.24). In the above three examples these would be $N^* \leftrightarrows N^y$, $N^* \leftarrow N^y$ and $N^* \rightarrow N^y$, respectively (albeit with bi-directed self-loops). However, in general it does not hold that the latent projection over eliminable nodes corresponds to the induced subgraph on the remaining nodes as bi-directed edges could occur between nodes within $\mathcal{N}_0$. Latent projections of causal graphs have been used to identify valid adjustment sets (Witte et al., 2020) to which we return in Section 7. We briefly comment on the projection graphs for the online supplementary material, examples (2, 3) in Supplement 1.

4.2 Identifiability and likelihood-ratio

Theorem 1 immediately implies that together with $\tilde{P} \ll P$ the likelihood-ratio

$$\frac{\mathrm{d}\tilde{P}|_{\mathcal{F}_t^{\mathcal{V}_0}}}{\mathrm{d}P|_{\mathcal{F}_t^{\mathcal{V}_0}}},$$

coincides with the weights of equation (9) with λ_t^* being the $\mathcal{F}^{\mathcal{V}_0}$-intensity of N^*. Hence, as a key implication of Theorem 1 we obtain that any causal parameter ξ that is a function only of $\tilde{P}|_{\mathcal{F}^{\mathcal{V}_0}}$ is identified from $\mathcal{F}^{\mathcal{V}_0}$ without requiring information on $\mathcal{U}$.

Continuing the above example with graph $U^1 \rightarrow N^* \leftrightarrows N^y \leftarrow U^2$: Assume N^* is a process indicating start of treatment and N^y a disease process, e.g. indicating a cardiovascular event. Then U^1 might be a process affecting the availability of the treatment but nothing else, e.g. a shortage in the pharmacy; U^2 might relate to events that affect the disease process but not the availability of, or decision to start, treatment, e.g. a change at the job. If we are interested in the effect of, say, early versus late start of treatment on cardiovascular problems, we wish to ignore U^1 and U^2. Then, assuming the local independencies implied by the δ-separations in the graph hold, and that G is causally valid, Theorem 1 tells us that the reduced system $N^* \leftrightarrows N^y$ is also causally valid and the likelihood-ratio is identified without information on (U^1, U^2). Hence we can identify the desired causal effect by reweighting according to (9) using the $\mathcal{F}^{N^*,N^y}$–intensity of N^* ignoring (U^1, U^2).

5 Identifiability under censoring

While the above deals with unmeasured processes, in time-to-event settings information is often incomplete due to right censoring because, e.g. follow-up time of studies is limited. The issue of

identifiability under right censoring can be seen from two subtly different angles: The approach via filtrations (Aalen et al., 2008) typically assumes there are processes and baseline variables of interest $\mathcal{V}_0$ and a separate censoring process N^c, where $\mathcal{F}_t^{\mathcal{V}_0 \cup N^c}$ is the filtration jointly generated by $\mathcal{V}_0 \cup \{N^c\}$, $\mathcal{F}_t^{\mathcal{V}_0}$ is generated by $\mathcal{V}_0$ alone, i.e. with no information on being censored, and $\mathcal{F}_{t\wedge C}^{\mathcal{V}_0 \cup N^c}$ being generated by the observable processes, i.e. with everything that is censored being 'invisible'. Identification is then about the possibility to use only information in $\mathcal{F}_{t\wedge C}^{\mathcal{V}_0 \cup N^c}$ to infer quantities such as intensities defined with respect to $\mathcal{F}_t^{\mathcal{V}_0}$. However, this presupposes that it is self-evident what real-life situation $\mathcal{F}_t^{\mathcal{V}_0}$ represents, i.e. how there can be no censoring. If censoring simply occurs due to the end of follow-up, then one can say that $\mathcal{F}_t^{\mathcal{V}_0}$ refers to the patients' health and lives regardless of whether they are being observed in a study or not, and hence regardless of end of follow-up. This point of view underlies the motivation of the assumption of *independent* censoring (Andersen, 2005; Andersen et al., 1993).

However, when other types of events, such as 'death from other causes', 'treatment switching', or even just 'drop-out' (often combined with a change in medical care) are considered as censoring events then it becomes less clear to what kind of situation the no-censoring filtration $\mathcal{F}^{\mathcal{V}_0}$ refers, let alone whether that filtration represents something practically meaningful. A different angle has therefore sometimes been adopted not only for survival analyses, but also in related problems such as missing data, drop-out or competing events (Farewell et al., 2017; Hernán & Robins, 2020; Young et al., 2020): Assume an overall model P for $\mathcal{V}$ including $\mathcal{V}_0$ and $\{N^c\}$; this is now modified to $\tilde{P}$ with λ^c replaced by $\tilde{\lambda}^c \equiv 0$ while all other local characteristics remain the same. In other words, we assume that a sufficiently rich system $\mathcal{V}$ can be conceived, such that the situation of interest, without censoring, can formally be described by a hypothetical intervention that sets the censoring intensity to zero within a causally valid system (cf. Definition 2). This might imply a much more fundamental change than simply hiding or revealing information. We believe that this approach based on 'preventing' censoring facilitates reasoning about the structural assumptions that allow identifiability of quantities in the uncensored situation, and would also highlight when censoring by certain types of events may not be practically meaningful. Even if censoring is simply an inability to observe the system but does not in itself affect the system, then violation of independent censoring can occur because censoring is *indirectly* informative for some hidden or ignored processes, and this can also easily be read off from local independence graphs thus alerting us to such a violation.

5.1 Independent censoring and causal validity

In this section, we link independent censoring to local independence, so that it can be read off from local independence graphs. Further, we consider a hypothetical intervention on the system to *prevent* censoring; hence we give further conditions for identifiability invoking causal validity.

Independent censoring is often used informally or confused with stochastic independence between processes. Here, following Andersen (2005), we formulate it in terms of local independence.

Definition 4 Let $\mathcal{P}$ be a local independence model with sets of variables or processes $A \subset \mathcal{N}$, $B \subset \mathcal{V}$, and N^c representing the counting process for censoring events. Then censoring is said to be independent for A, given B, if A is locally independent of N^c given $A \cup B$. In the special case where $A = \mathcal{N} \setminus \{N^c\}$ and $B = \mathcal{B}$, then we say that the whole model satisfies independent censoring.

Remark 4 As local independence can be read off from a local independence graph via δ-separation, the above can be checked graphically. Let $\mathcal{P}(G)$ be a local independence model on a graph G with sets of nodes $A \cup B \cup N^c \subset \mathcal{V}$. Censoring is independent for A, given $A \cup B$, if N^c is δ-separated from A by $A \cup B$ in G. The whole model on $\mathcal{V}$ satisfies independent censoring if the node N^c has no children in G. Moreover, the submodel induced by $A \cup B \cup \{N^c\}$ is subject to independent censoring if N^c is δ-separated from $A \cup B$ by $A \cup B$ alone in G (or, more explicitly, if $N^c \nrightarrow_G \mathcal{N} \cap (A \cup B) \mid A \cup B$).

In online supplementary material, Supplement 5 (Lemma 1) we prove that independent censoring ensures that the intensities with respect to the uncensored filtration are identifiable. We further argue that as long as there is a nonzero probability to observe the event before censoring (e.g. by the end of follow-up) there exists *de-censoring* maps ζ to obtain P on $\mathcal{F}_t^{\mathcal{V}_0}$ from information restricted to $\mathcal{F}_{t\wedge C}^{\mathcal{V}_0 \cup N^c}$ (see online supplementary material, Remark 1 in Supplement 5), i.e.

$$\zeta(P|_{\mathcal{F}_{T\wedge C}^{\mathcal{V}_0 \cup N^c}}) = P|_{\mathcal{F}_T^{\mathcal{V}_0}}. \tag{13}$$

The above formulation via the de-censoring map is very general; in practice, it is typical to show that a particular method of estimation is consistent for the desired parameter under independent censoring within a (semi-)parametric model. For our purposes, we choose to stay with the more general framework of identification just assuming the existence of intensities.

However, as can be seen from the example in online supplementary material, Supplement 6, reinterpreting N^A as censoring process, we note that the system (N^A, N^D) satisfies independent censoring but that this is not sufficient to ensure identifiability for a system in which censoring is *prevented* as it would yield the wrong intensity of N^D. We need the additional causal validity with regard to an intervention on censoring, as formalized next.

Proposition 2 Let $\mathcal{P}(G)$ be a local independence model on $\mathcal{V}$ subject to independent censoring with bounded intensity for censoring; additionally assume it is causally valid with respect to an intervention that *prevents* censoring.

If N^c is δ-separated from $\mathcal{N}_0$ given $\mathcal{V}_0$ in G, then the marginal model on $\mathcal{V}_0 \cup \{N^c\}$ satisfies independent censoring and retains causal validity with respect to the same intervention on N^c.

The proof is given in online supplementary material, Supplement 7. Put together, the above results mean that when we have independent censoring we can recover the uncensored $\mathcal{F}^{\mathcal{V}_0}$-intensities from censored data, and if additionally we have causal validity with regard to an intervention that prevents censoring then these are also the intensities under the interventional distribution $\tilde{P}$ where censoring is prevented. Note that the results are general in that they do not rely on any particular structure for the intensities other than that they exist.

5.2 'Randomizing' censoring

Here we discuss a slight generalization of the above that will make estimation more efficient. In fact, Proposition 2 remains true for an intervention on N^c that imposes a different $\mathcal{F}^{\mathcal{V}_0 \cup N^c}$-intensity for censoring (cf. the proof in the supplement); 'preventing' censoring is just a special case. We will loosely refer to censoring interventions that impose an $\mathcal{F}^{\mathcal{V}_0 \cup N^c}$-intensity as 'randomizing' censoring. This may not be an intervention that is in itself of practical relevance, but, as mentioned in Section 4.4, leads to more stable weights such as (14) in the following section. In other words, for reweighting we want to consider intervention intensities that are 'not too different' from the observational censoring intensity. The intervention that prevents censoring, corresponding to setting the censoring intensity to zero, may not yield efficient estimation. Formally, we posit the following.

Remark 5 We assume that if a model is causally valid with respect to an intervention preventing censoring, then it is also causally valid with respect to an intervention randomizing censoring.

The following corollary justifies that we can essentially equate the two types of interventions on censoring, as an intervention that randomizes censoring yields the same hypothetical distribution as one that prevents censoring when restricted to $\mathcal{F}^{\mathcal{V}_0}$.

Corollary 1 Let $\mathcal{P}(G)$ be as in Proposition 2 and assume that the model is causally valid with respect to prevention of censoring, with P^p being the model in which censoring is prevented. Let P^r be a model where we have imposed an

$\mathcal{F}^{\mathcal{V}_0 \cup N^c}$-intensity for censoring. Then, we have that

$$P^p|_{\mathcal{F}_t^{\mathcal{V}_0}} = P^r|_{\mathcal{F}_t^{\mathcal{V}_0}} = P|_{\mathcal{F}_t^{\mathcal{V}_0}}.$$

In particular, every parameter on such hypothetical measures restricted to $\mathcal{F}^{\mathcal{V}_0}$ is invariant with respect to the choice of censoring intervention.

Together with online supplementary material, Lemma 1 (Supplement 5) we have that the $\mathcal{F}^{\mathcal{V}_0}$-intensity of every $N \in \mathcal{V}_0$ is identifiable by the observable censored information *and retains its interpretation under a hypothetical scenario where censoring can be prevented*. In this case, we do not need to use reweighting to identify parameters under prevention of censoring. However, in the following Section 7 we consider the case where independent censoring does not necessarily hold with respect to $\mathcal{V}_0$ but may require further information, which can only be marginalized out after reweighting.

6 MSMs and censoring

In this section, we combine and further generalize the previous results. We wish to draw inference on the effect of a hypothetical intervention on a treatment or exposure process N^x on one or more outcome processes $\mathcal{N}_0$ under a further intervention that prevents censoring. The set $\mathcal{N}_0$ could include a survival-type outcome, but can be much more general event-histories such as recurrent events and multi-state processes. Typically, causal validity will not hold for these sets alone, and adjustment is required for additional covariates, baseline or processes, denoted $\mathcal{L}$. Therefore, the set of measured $\mathcal{V}_0$ from the previous sections is now extended to $\mathcal{V}_0 \cup \mathcal{L}$. Here $\mathcal{L}$ is not of substantive interest in the sense that we would like the effect of the treatment intervention on $\mathcal{N}_0$ marginally over $\mathcal{L}$. This can be regarded as a continuous-time event-history analogue to the causal parameter of marginal structural models (Hernán et al., 2000; Robins et al., 2000; Røysland, 2011). The set $\mathcal{L}$ is typically needed for adjustment if it contains processes that are not eliminable, e.g. time-dependent confounding. In other words, the local independencies of Theorem 1 and Proposition 2 may not hold for $\mathcal{V}_0$ alone but would hold for $\mathcal{V}_0 \cup \mathcal{L}$. As in discrete time, a continuous-time MSM can be fitted using suitable reweighting. However, the reweighting has two aspects: mimicking an intervention that prevents (or randomizes) censoring (cf. Remark 5), and an intervention on the treatment process.

Let the stochastic system be described by the following sets of variables and processes:

$$\mathcal{V} = \mathcal{V}_0 \dot\cup \mathcal{L} \dot\cup \mathcal{U} \dot\cup \{N^c\}$$

where $\mathcal{V}_0 = \mathcal{B}_0 \cup \mathcal{N}_0$, $N^x \in \mathcal{N}_0$ is a treatment process, N^c the censoring process, $\mathcal{N}_0^{\backslash x} = \mathcal{N}_0 \backslash \{N^x\}$ the outcome processes of interest, $\mathcal{L} = \mathcal{B}_\mathcal{L} \cup \mathcal{N}_\mathcal{L}$ are the measured baseline variables $\mathcal{B}_\mathcal{L}$ and counting processes $\mathcal{N}_\mathcal{L}$ we wish to marginalize out, and $\mathcal{U}$ unobserved variables or processes. The interventions replace the $\mathcal{F}^\mathcal{V}$ intensities of this system by new intensities for N^c and N^x. As we cannot observe $\mathcal{U}$, we now give conditions such that we can instead work with the observable intensities, where λ^c is the $\mathcal{F}^{\mathcal{V}_0 \cup \mathcal{L} \cup N^c}$-intensity of N^c with respect to P, while λ^x denotes the $\mathcal{F}^{\mathcal{V}_0 \cup \mathcal{L} \cup N^c}$-intensity of N^x with respect to P. In contrast, the interventions enforce an $\mathcal{F}^{\mathcal{V}_0 \cup N^c}$-predictable $\tilde\lambda^c := \rho^c \cdot \lambda^c$, and an $\mathcal{F}^{\mathcal{V}_0}$-predictable $\tilde\lambda^x := \rho^x \cdot \lambda^x$ as in Proposition 1. As Theorem 2, shows, we can obtain the hypothetical scenario from the observables under key structural assumptions using the following combined weights

$$W_t = \prod_{s \le t} (\rho_s^c)^{\Delta N_s^c} \exp\Big\{ - \int_0^t (\rho_s^c - 1)\lambda_s^c \mathrm{d}s \Big\} \prod_{s \le t} (\rho_s^x)^{\Delta N_s^x} \exp\Big\{ - \int_0^t (\rho_s^x - 1)\lambda_s^x \mathrm{d}s \Big\}. \tag{14}$$

The first part of the weight refers to the censoring reweighting, and is given by $\exp\{-\int_0^t (\rho_s^c - 1)\lambda_s^c \mathrm{d}s\}$ before censoring.

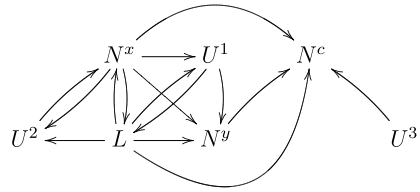


Figure 1. Illustration of Theorem 2.

Theorem 2 With the above notation and set-up, consider a local independence model $\mathcal{P}(G)$. We assume that censoring is independent with respect to $\mathcal{V}$, i.e. $\mathrm{ch}(N^c) = \emptyset$ in G. Furthermore, assume causal validity with respect to an intervention that prevents censoring and the additional intervention on the treatment process. Let $\tilde{\mathcal{P}}$ denote the resulting interventional model.

Further, assume

(i) $N^c \nrightarrow_G (\mathcal{N}_0, \mathcal{N}_\mathcal{L}) | (\mathcal{V}_0, \mathcal{L})$, and
(ii) $\mathcal{U}$ is eliminable with respect to $(N^x, \mathcal{L} \cup \mathcal{V}_0^{\setminus x})$;

finally, assume the technical conditions that ρ^c and ρ^x are bounded. Then we have:

The interventional distribution $\tilde{P}$ under both hypothetical interventions (preventing censoring and intervening on treatment) restricted to the subset $\mathcal{V}_0$ is identified from the observable information $\mathcal{F}_{T\wedge C}^{\mathcal{V}_0 \cup \mathcal{L} \cup N^c}$ by

$$\tilde{P}|_{\mathcal{F}_t^{\mathcal{V}_0}} = \zeta\left\{ \left(W_{t\wedge C} \cdot P|_{\mathcal{F}_{t\wedge C}^{\mathcal{V}_0 \cup \mathcal{L} \cup N^c}} \right) \Big|_{\mathcal{F}_{t\wedge C}^{\mathcal{V}_0 \cup N^c}} \right\}.$$

Thus, the marginal density $\tilde{P}|_{\mathcal{F}_t^{\mathcal{V}_0}}$ is given by reweighting, marginalizing over $\mathcal{L}$ and applying a decensoring map as in (13) (see online supplementary material, Remark 1 in Supplement 5).

The proof is given in online supplementary material, Supplement 8. In words, the above theorem states conditions such that any marginal (over $\mathcal{L}$) causal parameters are identified from the censored data ignoring $\mathcal{U}$, where these causal parameters quantify the effect on $\mathcal{N}_0$ of an intervention on N^x while preventing censoring. In particular, with Theorem 2(i), we have independent censoring for $\tilde{P}$ restricted to $\mathcal{V}_0 \cup \{N^c\}$. And Theorem 2(ii) can be seen as analogous to the sequential exchangability (or ignorability) assumption in discrete-time sequential treatment estimation. Due to the generality of Theorem 2, we can use the above reweighting strategy in any standard survival analysis methods to estimate parameters under $\tilde{P}|_{\mathcal{F}^{\mathcal{V}_0}}$ (see also Ryalen et al., 2019).

In Figure 1, we give a graphical example to illustrate Theorem 2. Here, all nodes are processes and $\mathcal{V}_0 = \{N^y, N^x\}$, $\mathcal{L} = \{L\}$, $\mathcal{U} = \{U^1, U^2, U^3\}$. The example represents a situation of time-dependent confounding by the process L: it is affected by, and affects itself the treatment process N^x while also affecting the outcome process N^y. We see that censoring is independent in $\mathcal{V}$ as the node N^c is childless. Property (i) can easily be seen via δ-separation; note that while N^c is locally dependent on U^3, the latter does not affect the remaining nodes, so that marginally over U^3 independent censoring is retained. Property (i) would be invalid if U^1 or U^2 had directed edges pointing at N^c. In G, the unobservable processes (U^1, U^2, U^3) are eliminable in any sequence. We can verify that (L, N^y) are locally independent of U^2 given (L, N^y, N^x, U^1, U^3), and N^x is locally independent of U^1 given (N^x, L, N^y, U^3); and further (L, N^y, N^x) are locally independent of U^3, so that property (ii) holds. Note that if we modified the example to L being unobservable so that $\mathcal{L} = \emptyset, \mathcal{U} = \{L, U^1, U^2, U^3\}$ then neither (i) nor (ii) of Theorem 2 would hold.

Theorem 2 can also be used in the following way (similar to Pearl & Robins, 1995 for discrete time): Let us partition the nodes into

$$\mathcal{V} = \mathcal{V}_0 \cup \mathcal{Z} \cup \{N^c\}.$$

where $\mathcal{Z}$ is any set of baseline variables or additional processes deemed relevant for causal validity. Then, if a subset $\mathcal{L} \subset \mathcal{Z}$ *exists* such that Theorem 2 holds with $\mathcal{U} = \mathcal{Z} \backslash \mathcal{L}$, then the interventional distribution $\tilde{P}$ is identified if $\mathcal{L}$ can be measured. We leave the development of algorithms that find such $\mathcal{L}$ for future work.

A slight generalization of Theorem 2 can be obtained: We considered the intensities λ^c and λ^x of N^c and N^x that were measurable with respect to the observed information $\mathcal{F}^{\mathcal{V}_0 \cup \mathcal{L} \cup N^c}$ and P. We could instead accommodate different adjustment sets in the sense of allowing λ^c and λ^x to be intensities with respect to smaller filtrations, say, $\mathcal{F}^{\mathcal{V}_0^c \cup \mathcal{L}^c \cup N^c}$ and $\mathcal{F}^{\mathcal{V}_0^x \cup \mathcal{L}^x \cup N^c}$, respectively, with $\mathcal{V}_0^c, \mathcal{V}_0^x \subset \mathcal{V}_0$ and $\mathcal{L}^c, \mathcal{L}^x \subset \mathcal{L}$: conditions (i) and (ii) still ensure the result. In particular, the likelihood ratio then still coincides with (14).

7 Application: introducing HPV-testing to follow-up low-grade cytology exams in cervical cancer screening programme in Norway

Cervical cancer is an infrequent end-stage of common cellular changes, starting with minor abnormalities and ranging through more definitely premalignant change to localized invasive and disseminated disease to death. This is an extremely complex process, but being able to detect cancer in its early stages or as precancers, accompanied by prompt appropriate treatment, are key elements of successful cancer screening programmes.

Since 1995, Norwegian women 25–69 years of age are advised to attend cervical cancer screening every three years for cytology exam, with the objective to identify and treat those with cervical intraepithelial lesion grade 2 or 3 (CIN2+). Some of the cytology exams yield inconclusive results, and since 2005 HPV testing has been used to guide future treatment strategies.

7.1 Which HPV-tests are suitable for secondary screening?

The three most common HPV tests in Norway from 2005 to 2010 were AMPLICOR HPV Test, Hybrid Capture2 High-Risk HPV DNA Test or PreTectTM HPV-Proofer referred to as Amplicor, HC2, and PreTectProofer (Nygård et al., 2014). When used after an inconclusive finding, PreTectProofer negative HPV-tests were more often followed later by a detection of CIN2+ than its competitors suggesting more false-negative tests for PreTecProofer (Haldorsen et al., 2011; Nygård et al., 2014). However, PreTectProofer patients were also subject to more subsequent testing (Nygård et al., 2014), presumably due to the manufacturer's recommendations. Thus, the apparent false-negative PreTectProofer results might have been due to the higher rate of subsequent testing.

The objective of our analysis is to compare the cumulative incidences of CIN2+ detection in the PreTectProofer group with the other two groups under a hypothetical scenario where an intervention ensures that the PreTectProofer patients are subject to the same rate of subsequent testing as under the other test-types. More formally, let $\tilde{P}$ denote the distribution under the modified 'subsequent testing' intensity and prevention of censoring. The contrast of interest is then given as

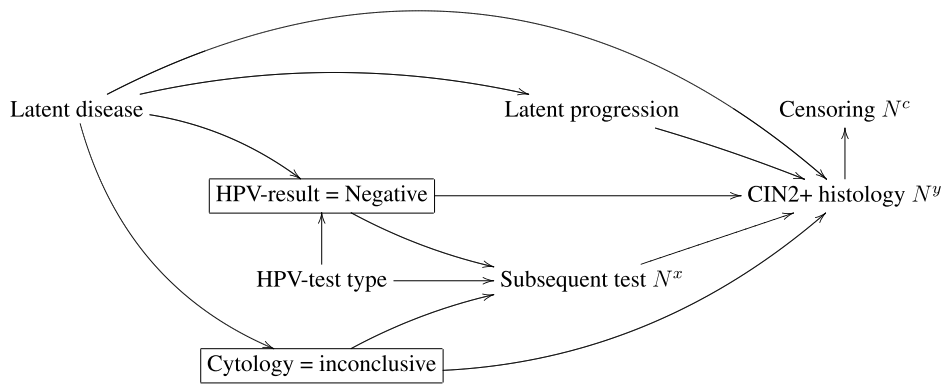


Figure 2. Local independence graph showing the assumed structure of the HPV-testing scenario.

the difference in cumulative incidence functions over time t,

$$\tilde{P}(N_t^y = 1 \mid \text{PTP, neg.HPV, inconcl.Cyt}) - \tilde{P}(N_t^y = 1 \mid \text{A/HC, neg.HPV, inconcl.Cyt}). \quad (15)$$

Under the assumed causal structure described next, this is not zero only if the different HPV-test types have different false-negative rates.

7.2 Local independencies and causal validity

Figure 2 contains a local independence model for the assumed HPV-testing scenario. Here we have that 'latent disease' represents a baseline variable describing the disease at the time of the HPV-test and cytology. 'HPV-test type' is the HPV-test that accompanies the cytology (binary: PreTectProofer, yes or no). The node 'latent progression' represents a set of counting processes describing the disease's progression from the time of the cytology over time. 'Subsequent test' is a counting process that 'counts' the first subsequent cytology or HPV-test. The node 'CIN2+ histology' represents the counting process that jumps when CIN2+ is detected. The analysis is restricted to subjects who had an inconclusive (i.e. ASCUS/LSIL/unsatisfactory) secondary cytology screening and who initially had a negative HPV-test result, as indicated by the boxed nodes in Figure 2. Individuals are censored at the end of the follow-up period if there was no occurrence of CIN2+ detection. The number of deaths was negligible, and is ignored.

Key assumptions are that any testing in itself does not affect the disease progression, but also vice versa, disease progression does not affect the testing regime. This could be violated if certain (undocumented) symptoms lead to the initiation of an HPV-test, but this is unlikely in the present case. The edge 'HPV-test type' → 'Subsequent test' is due to the observationally differing subsequent testing rates. In this analysis, we mimic an intervention imposing the same subsequent testing regime for all the HPV tests, in effect making 'Subsequent test' locally independent of 'HPV-test type' in the hypothetical scenario. Under $\tilde{P}$, any association between type of HPV-test and CIN2+ histology when conditioning on {'HPV-result'=negative, 'Cytology'=inconclusive} is then due to unblocked paths 'HPV-test type' → 'HPV-result' ←'Latent disease' → ⋯'CIN2+ histology' which would indicate a tendency to false-negative results due to the edge 'HPV-test type' → 'HPV-result'.

We will appeal to Theorem 2. For this we define the sets of nodes

$$\mathcal{B}_0 = \{\text{Cytology, HPV-test type, HPV result}\}$$
$$\mathcal{U} = \{\text{Latent disease, Latent progression}\}.$$

Let N^x, N^y, N^c be counting processes where N^x counts initiation of 'Subsequent test', N^y counts histology finding CIN2+, and N^c counts censoring events. Thus, $\mathcal{N}_0 = \{N^x, N^y\}$, and $\mathcal{N}_0^{\setminus x} = N^y$ and $\mathcal{V}_0 = \mathcal{B}_0 \cup \mathcal{N}_0$. We make the following observations:

- N^c has no descendants, i.e. censoring is independent with respect to $\mathcal{V}$.
- There are no allowed trails from N^c to $\mathcal{V}_0$, so $N^c \nrightarrow_G \mathcal{N}_0 | \mathcal{V}_0$, and condition (i) from Theorem 2 holds.
- Every allowed trail from $\mathcal{U}$ to N^x is blocked by either 'Cytology' or 'HPV-test type', both of which are in $\mathcal{B}_0$. Hence, $\mathcal{U}$ is eliminable with respect to $(N^x, \mathcal{V}_0 \setminus \{N^x\})$ in G, and condition (ii) of Theorem 2 holds.

These points justify the use of Theorem 2, and the $\mathcal{F}^{\mathcal{V}_0}$-intensity of N^y under $\tilde{P}$ is identified. As $\mathcal{L} = \emptyset$, we have from Proposition 2 that the model restricted to $\mathcal{F}^{\mathcal{V}_0 \cup N^c}$ is causally valid with respect to prevention of censoring and subject to independent censoring. The censoring weights thus equal one, and (14) reduces to $W_t = \prod_{s \le t} (\rho_s^x)^{\Delta N_s^x} \exp\{-\int_0^t (\rho_s^x - 1)\lambda_s^x \mathrm{d}s\}$ in this example.

7.3 Analysis

We consider data from the Cancer Registry of Norway on 1736 subjects (878 in the PreTectProofer group and 858 in the Amplicor/HC2 group) with inconclusive cytology and

negative initial HPV test recorded in 2005–2010 until CIN2+ or end of 2010 (for details see online supplementary material, Supplement Section 9). We calculate the probability of having CIN2+ detected by time t in a situation where individuals receive 'subsequent test' with intensity $\tilde{\lambda}^x$ equal to the intensity in the (pooled) Amplicor/HC2 group. The probability of interest is calculated by one minus the weighted Kaplan–Meier estimator $\hat{S}^w$, given by

$$\hat{S}^w_t = \prod_{T_i \le t} \left(1 - \frac{\widehat{W}^i_{T_i-} Y^i_{T_i}}{\sum_j \widehat{W}^j_{T_i-} Y^j_{T_i}} \right), \tag{16}$$

where $\widehat{W}^i$ are estimates of (14), the T_i's the observed detection times of CIN2+, and the Y^i_t's are at-risk indicators (not censored) at time t in a given group.

For the Amplicor/HC2 group, the 'subsequent test' intensity is equal to the observational intensity, and each W^i is equal to one. Thus, (16) reduces to the standard Kaplan–Meier estimator with CIN2+ occurrence as the endpoint.

To estimate this probability in the PreTectProofer-group, we first need estimates $\widehat{W}^i$. The $\mathcal{F}^{\mathcal{V}_0}$-intensity of N^x is only a function of 'HPV-test type' (due to local independences implied by the graph). We thus obtain the estimator

$$\widehat{W}^i_t = 1 + \int_0^t \widehat{W}^i_{s-}(\theta_{s-} - 1) dN^{x,i}_s - \int_0^t \widehat{W}^i_{s-} I(N^{x,i}_{s-} = 0) d(\hat{\hat{A}}^x_s - \hat{A}^x_s),$$

where $\theta_s = \frac{\hat{\hat{A}}^x_s - \hat{\hat{A}}^x_{s-b}}{\hat{A}^x_s - \hat{A}^x_{s-b}}$ for a smoothing parameter b, and $\hat{\hat{A}}^x_s$ and $\hat{A}^x_s$ are the Nelson-Aalen estimators for subsequent test initiation applied to the Amplicor/HC2 group and the PreTectProofer group, respectively. We calculate these weights using the `R` package `ahw`. The weighted Kaplan–Meier estimator (16) is consistent for a bandwidth parameter $b = b(n)$ depending on the sample size n with $b(n) \underset{n\to\infty}{\to} 0$ and $\sup_{s \le T} \tilde{\alpha}^x_s / \alpha^x_s < \infty$; see Ryalen et al. (2019, Theorems 1 & 2) and Ryalen, Stensrud and Røysland (2018, Theorem 1).

7.4 Results

The cumulative incidences of CIN2+ detection are shown in Figure 3. In the upper panel, we see three curves: for PreTectProofer without and with reweighting, and the nonPreTectProofer group without weighting. Thus, the proportion of CIN2+ detected in the PreTectProofer group is somewhat lower under the hypothetical subsequent-testing regime than observationally, i.e. without an intervention. However, the estimated contrast (15), shown in the lower panel, is still significant. Under the structural assumptions we made, this gives support to the interpretation that there are genuinely more false-negative HPV-test results in the PreTectProofer group than with the other test types, and that the greater CIN2+ numbers are not only a consequence of more frequent subsequent testing.

8 Conclusions and discussion

We proposed a formal graphical approach to causal reasoning in survival and general event-history settings in continuous time with graphical rules for the identifiability of (marginal) causal parameters. We formalized these in terms of interventions that modify the intensity of a treatment process which is similar to the notion of 'randomized plans' (Gill & Robins, 2001) or stochastic interventions (Dawid & Didelez, 2010; Díaz & van der Laan, 2018); and it is more explicit than that of 'causal influence' (Commenges & Gégout-Petit, 2009). Further, we conjecture that our change of treatment intensity could be thought of as a stochastic change of time: one can construct a stochastic time-change ϱ such that the P-intensity of N^x_ϱ coincides with the $\tilde{P}$-intensity of N^x (Andersen et al., 1993, II 5.2.2). While the predominant causal approach uses potential outcomes, we have chosen to simply compare the observational and the interventional distributions, P and $\tilde{P}$; this is similar in spirit to other causal frameworks, for instance by Spirtes et al. (2000), Dawid and Didelez (2010) and Peters et al. (2016).

Our criteria for identifiability are sufficient, and we believe that necessary conditions analogous to Shpitser and Pearl (2008) will be difficult to derive for the general continuous-time case. In future work, it will be interesting to generalize our results to the extended local independence graphs

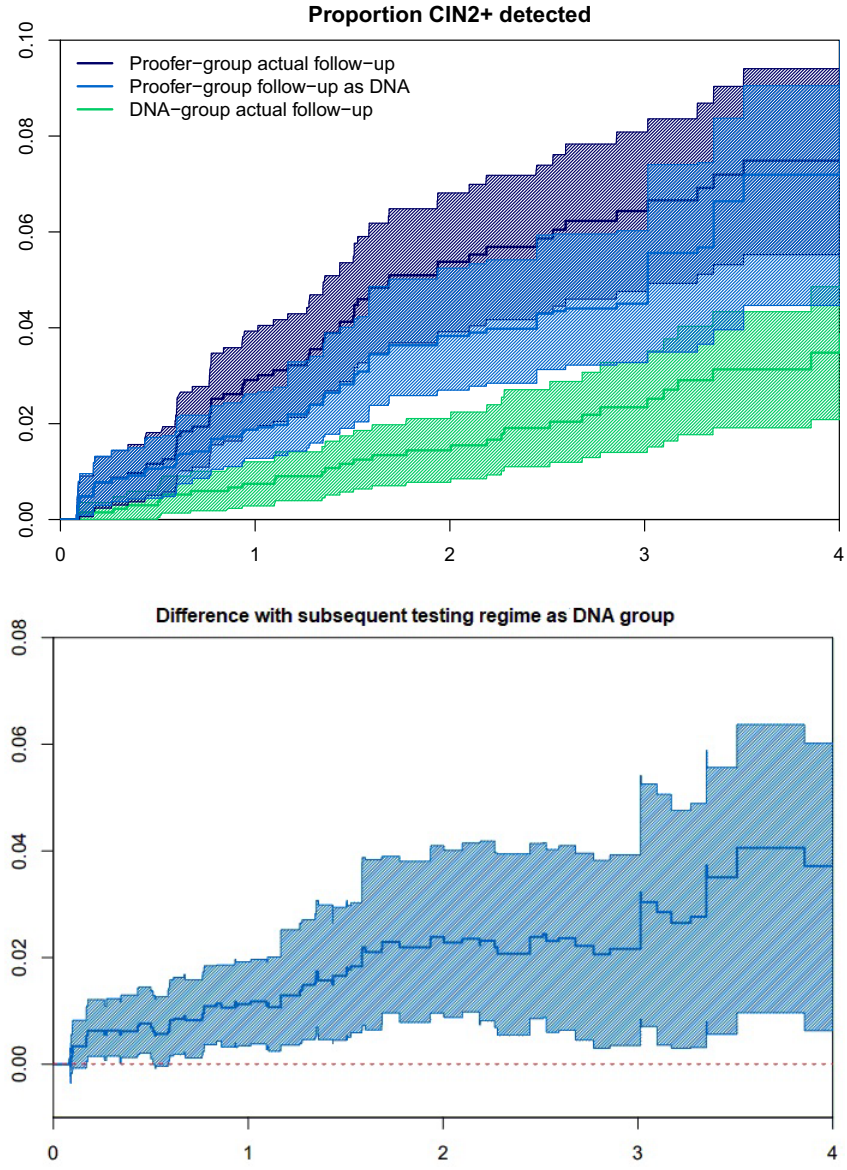


Figure 3. Upper: Proportion CIN2+ detected after the secondary screening with ASCUS/LSIL/unsatisfactory cytology and negative HPV-test. Lower: Difference between the proportions of detected CIN2+ in the PreTectProofer-group and Amplicor/HC2-group when imposing the Amplicor/HC2 group's subsequent testing regime on both groups. We obtained 95% pointwise confidence intervals using a bootstrap sample of 400.

of Mogensen and Hansen (2020) which are closed under marginalization and which can be obtained by projecting over unobservable processes. The requirement that processes may not jump simultaneously is plausible as long as the components truly represent separate phenomena, where accidental violations due to rounding are not essential. Alternatively, one can introduce additional processes which count simultaneous events, which will work in practice as long as the treatment and outcome processes do not jump systematically at the same time; redefining some of the other processes will not change the interpretation of the causal parameters of interest. If treatment or outcome processes are affected, future work would need to explicitly integrate systematic violations, which will require a more complex likelihood ratio process. Extending graph separation and eliminability to this situation will also require notable technical efforts outside the

scope of our paper. Extensions to more general processes would also be desirable; some such extensions of the graphical representation exist for larger classes of stochastic differential equations (Mogensen & Hansen, 2022; Mogensen et al., 2018) and stochastic kinetic models (Bowsher, 2010). Other future generalizations might allow for processes that can jump with continuous magnitude. Carrying out such extensions requires consideration of more general counting measures than we studied here.

We further addressed independent censoring arguing that inference for the uncensored case additionally requires causal validity with regard to an intervention that prevents censoring. While this has been recognized for longitudinal settings with drop-out (Hernán & Robins, 2020), it seems less appreciated in more traditional approaches to survival analysis; an exception is recent work by Rytgaard et al. (2021). Our results on identifiability under censoring are more widely applicable even outside a causal inference context, for instance to determine a sufficient set of covariates for IPCW.

Finally, Theorem 2 enables identification of causal parameters via reweighting, thus generalizing well-known results from discrete-time MSMs (Hernán et al., 2000; Robins et al., 2000). MSMs are often linked to the problem of time-dependent confounding but can also be used in other scenarios (Joffe et al., 2004). In the application in Section 8, Theorem 2 was used to establish that time-dependent confounding was not an issue (as $\mathcal{L}$ was the empty set) and that a fairly simple weighting process sufficed. Similarly, Ryalen, Stensrud, Fosså et al. (2018) used a continuous-time MSM to compare the treatment regimens radiotherapy and radical prostatectomy.

Our results apply to general multivariate counting processes, which include, e.g. multi-state processes. In particular, they do not rely on any particular (semi)-parametric class of models. While most practical inference needs additional modelling assumptions, the data example of Section 8 allowed for nonparametric estimation. In addressing identifiability, we have chosen the reweighting route which appears natural in view of the simplicity of Proposition 1 and corresponds to a change of measure technique. In discrete-time settings, g-computation is an alternative, or doubly robust and machine-learning extensions thereof (Kallus & Uehara, 2022; Luckett et al., 2020; Nie et al., 2020; Zhang et al., 2013). However, g-computation seems hard in entirely general continuous-time settings, as discussed by Gill (2001) (see also Gill & Robins, 2001), but fully parametric versions exist (Gran et al., 2015). We believe that our graphical causal reasoning can also be combined with g-estimation (Lok et al., 2004; Lok, 2008), or targeted minimum-loss estimation (Rytgaard et al., 2021) in continuous-time settings. It complements these methods because the graphical representation and explicit discussion of eliminability strengthens the plausibility of assumptions, such as sequential (conditional) exchangeability.

Conflict of interest: None declared.

Data availability

We do not have the rights to share the HPV dataset. The data can be obtained upon application at the Cancer Registry of Norway. The `R` implementation of the analysis in Section 8.3, which has been applied to a simulated dataset, is available on the GitHub repository `github.com/palryalen/paper-code`.

The simulated dataset, `sim_data.RData`, has identical features to the real dataset.

Supplementary material

Supplementary material is available online at *Journal of the Royal Statistical Society: Series B*.

References

Aalen O., Borgan O., & Gjessing H. (2008). *Survival and event history analysis: A process point of view*. Springer-Verlag.

Aalen O., Røysland K., Gran J., & Ledergerber B. (2012). Causality, mediation and time: A dynamic viewpoint. *Journal of the Royal Statistical Society: Series A*, *175*(4), 831–861. https://doi.org/10.1111/j.1467-985X.2011.01030.x

Aalen O. O. (1987). Dynamic modelling and causality. *Scandinavian Actuarial Journal*, *1987*(3-4), 177–190. https://doi.org/10.1080/03461238.1987.10413826

Andersen P. K. (2005). Censored data. In P. Armitage, & T. Colton (Eds.), *Encyclopedia of biostatistics*.
Andersen P. K., Borgan Ø., Gill R. D., & Keiding N. (1993). *Statistical models based on counting processes*. Springer series in statistics. Springer-Verlag.
Arjas E., & Parner J. (2004). Causal reasoning from longitudinal data. *Scandinavian Journal of Statistics*, *31*(2), 171–187. https://doi.org/10.1111/sjos.2004.31.issue-2
Bowsher C. G. (2010). Stochastic kinetic models: Dynamic independence, modularity and graphs. *The Annals of Statistics*, *38*(4), 2242–2281. https://doi.org/10.1214/09-AOS779
Commenges D., & Gégout-Petit A. (2009). A general dynamical statistical model with causal interpretation. *Journal of the Royal Statistical Society: Series B*, *71*(3), 719–736. https://doi.org/10.1111/j.1467-9868.2009.00703.x
Dawid A. P. (2002). Influence diagrams for causal modelling and inference. *International Statistical Review*, *70*(2), 161–189. https://doi.org/10.1111/insr.2002.70.issue-2
Dawid A. P., & Didelez V. (2010). Identifying the consequences of dynamic treatment strategies: A decision-theoretic overview. *Statistics Surveys*, *4*, 184–231. https://doi.org/10.1214/10-SS081
Díaz I., & van der Laan M. J. (2018). *Stochastic treatment regimes* (pp. 219–232). Springer International Publishing.
Didelez V. (2006). Asymmetric separation for local independence graphs. In *Proceedings of the 22nd conference in uncertainty in artificial intelligence*. AUAI Press.
Didelez V. (2007). Graphical models for composable finite Markov processes. *Scandinavian Journal of Statistics*, *34*(1), 169–185. https://doi.org/10.1111/sjos.2007.34.issue-1
Didelez V. (2008). Graphical models for marked point processes based on local independence. *Journal of the Royal Statistical Society: Series B*, *70*(1), 245–264. https://doi.org/10.1111/j.1467-9868.2007.00634.x
Didelez, V. (2018). Causal concepts and graphical models. In M. Maathuis, M. Drton, S. Lauritzen, & M. Wainwright (Eds.), *Handbook of graphical models* (pp. 353–380). CRC Press.
Farewell D. M., Huang C., & Didelez V. (2017). Ignorability for general longitudinal data. *Biometrika*, *104*(2), 317–326. https://doi.org/10.1093/biomet/asx020
Gill R. D. (2001). Causal inference for complex longitudinal data: The continuous time g-computation formula.
Gill R. D., & Robins J. M. (2001). Causal inference for complex longitudinal data: The continuous case. *Annals of Statistics*, *29*(6), 1785–1811. https://doi.org/10.1214/aos/1015345962
Gran J. M., Lie S., Øyeflaten I., Borgan Ø., & Aalen O. O. (2015). Causal inference in multi-state models—sickness absence and work for 1145 participants after work rehabilitation. *BMC Public Health*, *15*(1), 1–16. https://doi.org/10.1186/s12889-015-2408-8
Haldorsen T., Skare G. B., & Bjørge T. (2011). Sekundærscreening med HPV-tester i masseundersøkelsen mot livmorhalskreft. *Report from the Cancer Registry of Norway*.
Hansen N., & Sokol A. (2014). Causal interpretation of stochastic differential equations. *Electronic Journal of Probability*, *19*, 1–24. https://doi.org/10.1214/EJP.v19-2891
Hernán M., Brumback B., & Robins J. (2000). Marginal structural models to estimate the causal effect of zidovudine on the survival of HIV-positive men. *Epidemiology*, *11*(5), 561–570. https://doi.org/10.1097/00001648-200009000-00012
Hernán M., & Robins J. (2020). *Causal inference: What if*. Chapman & Hall/CRC.
Jacobsen, M. (2006). *Point process theory and applications. Marked point and piecewise deterministic processes* (1). Birkhäuser Boston. Series: Probability and Its Applications. https://doi.org/10.1007/0-8176-4463-6
Jacod J., & Shiryaev A. N. (2003). *Limit theorems for stochastic processes, vol. 288 of Grundlehren der Mathematischen Wissenschaften [fundamental principles of mathematical sciences]* (2nd ed.). Springer-Verlag.
Joffe M. M., Ten Have T. R., Feldman H. I., & Kimmel S. E. (2004). Model selection, confounder control, and marginal structural models: Review and new applications. *The American Statistician*, *58*(4), 272–279. https://doi.org/10.1198/000313004X5824
Kallsen J., & Shiryaev A. N. (2002). The cumulant process and Esscher's change of measure. *Finance and Stochastics*, *6*(4), 397–428. https://doi.org/10.1007/s007800200069
Kallus N., & Uehara M. (2022). Efficiently breaking the curse of horizon in off-policy evaluation with double reinforcement learning. *Operations Research*, *70*(6), 3282–3302. https://doi.org/10.1287/opre.2021.2249
Kennedy E. H. (2019). Nonparametric causal effects based on incremental propensity score interventions. *Journal of the American Statistical Association*, *114*(526), 645–656. https://doi.org/10.1080/01621459.2017.1422737
Last G., & Brandt A. (1995). *Marked point processes on the real line*. Probability and its applications (1995 ed.). Springer.
Lauritzen S. L. (1996). *Graphical models*. Oxford University Press.
Lok J., Gill R., Van Der Vaart A., & Robins J. (2004). Estimating the causal effect of a time-varying treatment on time-to-event using structural nested failure time models. *Statistica Neerlandica*, *58*(3), 271–295. https://doi.org/10.1111/stan.2004.58.issue-3

Lok J. J. (2008). Statistical modeling of causal effects in continuous time. *The Annals of Statistics*, *36*(3), 1464–1507. https://doi.org/10.1214/009053607000000820
Luckett D. J., Laber E. B., Kahkoska A. R., Maahs D. M., Mayer-Davis E., & Kosorok M. R. (2020). Estimating dynamic treatment regimes in mobile health using v-learning. *Journal of the American Statistical Association*, *115*(530), 692–706. https://doi.org/10.1080/01621459.2018.1537919
Manski C. F. (2003). *Partial identification of probability distributions*. Springer Science & Business Media.
Meek C. (1995). Strong completeness and faithfulness in Bayesian networks. In *Proceedings of the 11th conference on uncertainty in artificial intelligence* (pp. 411–418). Morgan Kaufmann Publishers Inc.
Mogensen S. W., & Hansen N. R. (2020). Markov equivalence of marginalized local independence graphs. *The Annals of Statistics*, *48*(1), 539–559. https://doi.org/10.1214/19-AOS1821
Mogensen S. W., & Hansen N. R. (2022). Graphical modeling of stochastic processes driven by correlated noise. *Bernoulli*, *28*(4), 3023–3050. https://doi.org/10.3150/21-BEJ1446
Mogensen S. W., Malinsky D., & Hansen N. R. (2018). Causal learning for partially observed stochastic dynamical systems. In A. Globerson & R. Silva (Eds.). *Proceedings of the 34th conference on uncertainty in artificial intelligence* (pp. 350–360). AUAI Press.
Nie X., Brunskill E., & Wager S. (2020). Learning when-to-treat policies. *Journal of the American Statistical Association*, *116*(533), 392–409. https://doi.org/10.1080/01621459.2020.1831925
Nygård M., Røysland K., Campbell S., & Dillner J. (2014). Comparative effectiveness of human papillomavirus testing in the cervical cancer screening programme in Norway. *BMJ Open*, *4*(1), e003460. https://doi.org/10.1136/bmjopen-2013-003460
Pearl J. (1995). Causal diagrams for empirical research. *Biometrika*, *82*(4), 669–688. https://doi.org/10.1093/biomet/82.4.669
Pearl J. (2009). *Causality—models, reasoning, and inference* (2nd ed.). Cambridge University Press.
Pearl J., & Robins J. (1995). Probabilistic evaluation of sequential plans from causal models with hidden variables. In *Proceedings of the eleventh annual conference on Uncertainty in Artificial Intelligence (UAI-95)* (pp. 444–453). Morgan Kaufmann.
Peters J., Bühlmann P., & Meinshausen N. (2016). Causal inference by using invariant prediction: Identification and confidence intervals. *Journal of the Royal Statistical Society, Series B*, *78*(5), 947–1012. https://doi.org/10.1111/rssb.12167
Robins J., Hernán M., & Brumback B. (2000). Marginal structural models and causal inference in epidemiology. *Epidemiology*, *11*(5), 550–560. https://doi.org/10.1097/00001648-200009000-00011
Robins J. M. (1986). A new approach to causal inference in mortality studies with sustained exposure periods—Application to control for the healthy worker survivor effect. *Mathematical Modelling*, *7*(9–12), 1393–1512. https://doi.org/10.1016/0270-0255(86)90088-6
Robins J. M. (2001). Data, design and background knowledge in etiologic inference. *Epidemiology*, *12*(3), 313–320. https://doi.org/10.1097/00001648-200105000-00011
Røysland K. (2011). A martingale approach to continuous time marginal structural models. *Bernoulli*, *17*, 895–915. https://doi.org/10.3150/10-BEJ303
Røysland K. (2012). Counterfactual analyses with graphical models based on local independence. *Annals of Statistics*, *40*(4), 2162–2194. https://doi.org/10.1214/12-AOS1031
Rubin D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, *66*(5), 688–701. https://doi.org/10.1037/h0037350
Ryalen P. C., Stensrud M. J., Fosså S., & Røysland K. (2018). Causal inference in continuous time: An example on prostate cancer therapy. *Biostatistics*, *21*(1), 172–185. https://doi.org/10.1093/biostatistics/kxy036
Ryalen P. C., Stensrud M. J., & Røysland K. (2018). Transforming cumulative hazard estimates. *Biometrika*, *105*(4), 905–916. https://doi.org/10.1093/biomet/asy035
Ryalen P. C., Stensrud M. J., & Røysland K. (2019). The additive hazard estimator is consistent for continuous time marginal structural models. *Lifetime Data Analysis*, *25*(4), 611–638. https://doi.org/10.1007/s10985-019-09468-y
Rytgaard H. C., Gerds T. A., & van der Laan M. J. (2021). Continuous-time targeted minimum loss-based estimation of intervention-specific mean outcomes.
Schweder T. (1970). Composable Markov processes. *Journal of Applied Probability*, *7*(2), 400–410. https://doi.org/10.2307/3211973
Shpitser I., & Pearl J. (2006). Identification of conditional interventional distributions. In *Proceedings of the 22nd conference in uncertainty in artificial intelligence*. AUAI Press.
Shpitser I., & Pearl J. (2008). Complete identification methods for the causal hierarchy. *Journal of Machine Learning Research*, *9*, 1941–1979. http://jmlr.org/papers/v9/shpitser08a.html
Spirtes P., Glymour C., & Scheines R. (2000). *Causation, prediction and search* (2nd ed.). MIT Press.
VanderWeele T. J., & Shpitser I. (2011). A new criterion for confounder selection. *Biometrics*, *67*(4), 1406–1413. https://doi.org/10.1111/biom.2011.67.issue-4

Witte J., & Didelez V. (2019). Covariate selection strategies for causal inference: Classification and comparison. *Biometrical Journal*, *61*(5), 1270–1289. https://doi.org/10.1002/bimj.v61.5

Witte J., Henckel L., Maathuis M. H., & Didelez V. (2020). On efficient adjustment in causal graphs. *Journal of Machine Learning Research*, *21*, 1–45. http://jmlr.org/papers/v21/20-175.html

Young J. G., Stensrud M. J., Tchetgen E. J. T., & Hernán M. A. (2020). A causal framework for classical statistical estimands in failure-time settings with competing events. *Statistics in Medicine*, *39*(8), 1199–1236. https://doi.org/10.1002/sim.v39.8

Zhang B., Tsiatis A. A., Laber E. B., & Davidian M. (2013). Robust estimation of optimal dynamic treatment regimes for sequential treatment decisions. *Biometrika*, *100*(3), 681–694. https://doi.org/10.1093/biomet/ast014