Causal reasoning from longitudinal data

Jan Parner
Department of Biostatistics
University of Copenhagen
Blegdamsvej 3
DK-2200 Copenhagen N
Denmark
J.Parner@biostat.ku.dk

Elja Arjas
Rolf Nevanlinna Institute

Research Reports A27
September 1999

Rolf Nevanlinna Institute
B.O.Box 4 (Yliopistonkatu 5)
FIN-00014 University of Helsinki
Finland

ISBN 952-9528-53-1
ISSN 0787-8338
YLIOPISTOPAINO

Summary

In this paper we present a general framework for causal reasoning in clinical trials and observational studies involving an embedded assignment mechanism. Using a marked point process approach we provide a true time continuous framework in which we state the assumption needed to allow causal reasoning in likelihood based statistical inference.

The framework is solely based on observed quantities and does not rely on the hypothesis of the existence of a unit causal effect defined upon counterfactual (potential) variables. It does, however, contain special cases that from a probabilistic point of view are indistinguishable from the Rubin causal model (Rubin, 1974, 1978) extended further by Robins (1986). The connections to graphical models (Pearl, 1995) and Granger causality (Granger, 1969) are discussed briefly as well as the connection to predictive distributions (Arjas and Eerola, 1993).

Keywords: marked point process; conditional independence; no unmeasured confounders; frequentist inference; Bayesian inference; predictive distributions

1. Background

A framework for causal analysis that has gained considerable popularity in the field of medical statistics is sometimes called the *Rubin causal model* (Rubin, 1974, 1978; Holland, 1986), based upon the ideas of Neyman (Neyman, 1923). Although Neyman's inspiration came from agriculture, in an attempt to assess the effect on yield outcome caused by the differences in sorts sowed on different plots, his formulation is very close to the causal model proposed by Rubin. Neyman's idea was to regard the potential yield outcome Y_{ik} when assigning a particular variety k on a given plot i as a priori fixed, with the only randomness in the experiment arising from the differences in soil conditions over plots. In particular, the rule for assigning varieties to plots was considered non-random. On each plot (from now on referred to as units) he defined the unit causal effect of applying variety k rather than variety l on plot i as the difference of potential outcomes, $Y_{i,k} - Y_{i,l}$. What Holland (1986) calls the *Fundamental Problem of Causal Inference* is the fact that only one variety can be applied on a particular plot at a time, thus making the unit causal effect unidentifiable from the observed data. Instead of imposing a statistical model that could lead to some identifiability under restrictive assumptions he chose to focus on the average (over plots) causal effect which is estimable from the data. Although the average causal effect is just a summary measure for a probably much more complex underlying causal process (think about the dependence of rain, sun and temperature), Neyman's formulation is conceptually made without reference to any probability model for the data.

One thing that apparently was not clear to Neyman was how to assign the different varieties to plots in order to balance unmeasured soil conditions over varieties. Fisher (1925) recommended that physical randomisation be made an integral part of the experimentation, although the reason for balancing observed as well as unobserved plot characteristics by this device was provided only later in Fisher (1935).

As realized by Cox (1958), the Neyman framework had the implicit assumption of *no interference between units*, which was later given the name *Stabile Unit Treatment Value Assumption* or SUTVA (Rosenbaum and Rubin, 1983).

A more general formulation of Neyman's ideas was given by Rubin (1978), who pointed out the conceptual importance of regarding the treatment assignment as an outcome of a stochastic process. By understanding this process, called the *assignment mechanism*, it is possible to weaken the assumption of randomisation from stratified randomised trials to more complicated treatment selection rules. The key assumption for valid causal inference, in

the Rubin causal model for unbiased estimation of the average causal effect, is the *ignorable treatment* assumption (Rubin, 1978) or *strongly ignorable treatment* assumption (Rosenbaum and Rubin, 1983). In a two-armed clinical trial in which patients are assigned to either placebo ($A = 0$) or active treatment ($A = 1$) according to the indicator A, this assumption is stated as

$$P(A = a \mid X, (Y_0, Y_1)) = P(A = a \mid X), \quad a = 0, 1$$

whenever $P(A = a \mid X) > 0$, where X denotes pretreatment measured covariates (e.g. age, gender, height) and Y_0 and Y_1 are the potential outcomes. In other words, A is assumed to be independent of (Y_0, Y_1) given X. Informally, this is the assumption that the only information carried in the assignment mechanism about future outcomes under placebo and treatment, respectively, comes through the covariates X.

One important implication of separating the treatment assignment from the rest of the experiment has been called *the bridge to observational studies* (see e.g. Rubin, 1991). This refers to the property that the average causal effects as defined by Neyman are estimable in observational studies when $\mathcal{L}(A \mid X)$ is known or recoverable from data and the assumption of strongly ignorable treatment assignment holds.

Although the presentation of the Rubin causal model is here limited to the dichotomous treatment case, our framework is easily modified to the case where A is discrete. When A is continuous, which is the case if A is drug dosage, the problem becomes more complicated as noted by Robins (1997) unless treatment is polytomised.

Important work by Robins (1986, 1997), Robins *et al.* (1992) has turned the Rubin/Neyman framework into a sequential design, but a time-continuous version of a real longitudinal study is only tentatively described in Robins (1998).

Besides of elucidating the ignorable treatment assumption by changing the terminology to *no unmeasured confounders*, one of Robins' many contributions has been to formulate this assumption in the case of a sequential trial design. Since this idea motivates our definition of no unmeasured confounders, we outline his work here in a simple setting.

In this sequential trial all individuals enter at time zero and are followed until time $T \in \mathbb{N}$. At each time point $t = 0, 1, \ldots, T-1$ covariate measurements X_t are immediately followed by a treatment decision A_t ($\in \{0, 1\}$) that remains the same until time $t+1$. At time T the response Y is measured. Let $a = (a_0, \ldots, a_{T-1})$ be the non-random treatment regime where treatment a_t is assigned at time point t and let Y_a be the potential outcome of Y if this specific treatment regime a was assigned. Hence, $P(Y_a \leq y) = P(Y \leq y \mid A = a)$

with $A = (A_0, \ldots, A_{T-1})$. As in the one-dimensional case, only one of the Y_a's corresponding to the actual received treatment regime is observed. Individual causal effects are defined as contrasts between Y_a and $Y_{a'}$ for suitable choices of a and a'.

As in the Rubin causal model, valid causal inference is only possible under the assumption of sequentially ignorable treatment assignment, or using Robins' terminology, when there are no unmeasured confounders. Formally, this is stated as independence between the potential outcome Y_a and the treatment assignment A_t about to be conducted, conditionally on all previously given treatments and all measured covariates up to and including time t:

$$Y_a \perp\!\!\!\perp A_t \mid (X_0, \ldots, X_t), (A_0, \ldots, A_{t-1}), \ \forall a, \forall t = 0, \ldots, T-1$$

whenever

$$P(A_t = a_t \mid X_0, \ldots, X_t, A_0, \ldots, A_{t-1}) > 0$$

(that is, only if the treatment is actually possible).

2. The marked point process approach

In clinical research and epidemiology, considering causal effects is often closely connected to the idea of a potentially manipulable action, such as treatment or exposure, that is influencing a response measure of interest. The simplest example of such study design is a clinical trial where the action is assignment to some particular treatment. More complicated designs arise, e.g., when considering the effect of smoking habits on coronary heart disease since, even though smoking itself is in principle a manipulable action, a decision to change such habits is not within the control of the investigator in an observational study.

Although Neyman's potential outcomes are conceptually appealing when discussing causal reasoning, they are not necessary for the framework presented here. The same is true for the unit causal effect defined upon the potential outcomes in that, even if it existed, it would be unidentifiable from the data. Whether or not the existence of a unit causal effect is necessary for the formulation of causality is subject to an ongoing discussion amongst philosophers.

We shall not here contribute to this discussion but rather investigate which assumptions appear to be necessary for drawing causal conclusions from statistical data analysis. As will be explained below, causal reasoning relies only on an assumption of conditional independence, and so, effects that

may have a causal interpretation can be quantified in more direct ways than by using only expected values and other scalar valued summary measures.

2.1. Framework

We shall here only consider designs involving one or more consecutive actions over time, followed by a measure of a response. Suppose that on a single individual i, a random number n_i events occur over time. At each event time T_{ik}, covariates X_{ik} are measured and an action A_{ik} follows immediately upon this. Hence, data on a single individual consist of $(T_{ik}, (X_{ik}, A_{ik}))$, $k = 1, \ldots, n_i$, with $0 \equiv T_{i0} < T_{i1} < T_{i2} < \ldots$ and a response Y_i which is not necessarily measured on every individual. The sampling scheme can be formulated naturally by use of marked point processes (see Brémaud (1981), Arjas (1989), Karr (1991), Andersen *et al.* (1993)), with T_{ik} being the event times and $Z_{ik} = (X_{ik}, A_{ik})$ being the corresponding marks. Whether the continuous time process $(Z_i(t))_{t\geq 0}$ generating Z_{ik} at time T_{ik} is continuous itself or not is not important, only that $Z_{ik} = Z_i(T_{ik})$.

Let $(\Omega, \mathcal{F})$ be a measurable space on which the marked point process $(T_{in}, Z_{in})_{n\geq 1}$ is defined and assume that the marks Z_{in} take their values in a space $(E_i, \mathcal{E}_i)$ where E_i is a Borel subset of a compact metric space and $\mathcal{E}_i$ the Borel σ-algebra of E_i. We refer to E_i as the mark space where elements are of the form (x, a).

In some designs, e.g. randomised trials, there may not be a covariate measurement preceding the action. The mark is then denoted by $(\varnothing, a)$. In other sampling schemes a number of repeated covariate measurements are performed before considering what action to take. This is for instance the case in the treatment of asthmatics where a daily recording of pulmonary function like FEV_1-measurements (Forced Expiratory Volume in one second) over a predetermined period of time is the basis for a treatment decision. In these cases the mark takes a value of the form $(x, \varnothing)$. Such marks may also arise in observational studies where the status of action is sampled less frequently than covariates, e.g., due to costs or ethical reasons. As a consequence we can write the mark space in the form $E_i = (E_{X_i} \cup \{\varnothing\}) \times (E_{A_i} \cup \{\varnothing\}) \setminus \{(\varnothing, \varnothing)\}$ where E_{X_i} is the range of all covariate measurements and E_{A_i} is the range of all possible actions.

To each $A \in \mathcal{E}_i$ we associate the individual counting process $N_{i,t}(A)$ defined by

$$N_{i,t}(A) = \sum_{n\geq 1} 1_{(Z_{in}\in A)}\, 1_{(T_{in}\leq t)}$$

and in particular the basic counting process is given as $N_{i,t} = N_{i,t}(E_i)$.

If the follow-up concerns m different individuals, we can from the individual processes construct a new marked point process $(T_n, Z_n)_{n\geq 1}$ with mark space $(E, \mathcal{E}) = (E_1 \times \cdots \times E_m,\ \mathcal{E}_1 \otimes \cdots \otimes \mathcal{E}_m)$ describing the whole sample. Here the sequence (T_n) is defined as the union of event times $T_{i,k}$, $1 \leq i \leq m, 1 \leq k \leq n_i$ and E_i refers to events that can happen to individual i. As a convention we can include the mark $\varnothing$ in each E_i, signifying that "nothing is recorded to happen to individual i" but then delete the element $(\varnothing, \varnothing, \ldots, \varnothing)$ from E. (Note that this product space construction is not multivariate in the sense that we would consider a vector process of the form $(N_{1,t}(A_1), \ldots, N_{m,t}(A_m))_{t\geq 0},\ A_i \in \mathcal{E}_i$, with the individual marked point processes as components and restricted in the way that no two individual processes are allowed to jump simultaneously.) To each $A \in \mathcal{E}$ we associate the counting process $N_t(A)$ defined by

$$N_t(A) = \sum_{n\geq 1} 1_{(Z_n \in A)}\, 1_{(T_n \leq t)}$$

and the basic counting process is given as $N_t = N_t(E)$.

Here we consider histories $\mathcal{F}_t$ satisfying

$$\mathcal{F} \supset \mathcal{F}_t \supset \tilde{\mathcal{F}}_t, \quad t \geq 0$$

where $\tilde{\mathcal{F}}_t$ is the internal pre-t history of $(T_n, Z_n)_{n\geq 1}$.

Let P be a probability measure defined on $(\Omega, \mathcal{F})$ and assume from now on that (N_t) is a P-nonexplosive process (i.e. $N_t < \infty$ P-a.s., $\forall t > 0$). Often covariate measurements and actions follow a predetermined protocol in which case (T_n) is a deterministic sequence of time points. If so, the counting process $N_t(A)$ does not have a $(P, \mathcal{F}_t)$-intensity kernel, which is the reason why we base the formulation below on compensators, or dual predictable projections (Jacod, 1975), rather than on the intensity process. By doing that, the framework will embed the discrete time model as wellAssume that the filtered probability space $(\Omega, \mathcal{F}, \mathcal{F}_t, P)$ satisfies "the usual conditions". The counting process $N_t(A)$ has then a unique $(P, \mathcal{F}_t)$-compensator $\Lambda_t(A)$ with local description of the form

$$d\Lambda_t(dz) = d\Lambda_t\, \Phi_t(dz)$$

To avoid further technicalities assume that the compensator $\Lambda_t = \Lambda_t(E)$ does not have a continuous singular part and hence each $\Lambda_t(A)$ admits a representation of the form

$$\Lambda_t(A) = \int_0^t \lambda_s(A)\, ds + \sum_{s\leq t} \Delta\Lambda_s(A) \tag{1}$$

Here $\Delta\Lambda_s(A) = \Lambda_s(A) - \Lambda_{s-}(A)$, and $\lambda_s(A)$ is the $(P, \mathcal{F}_t)$-intensity of $N_t(A)$. Since (N_t) is P-nonexplosive $\Delta\Lambda_s(E)$ is non-zero for only a countable number of time points s. The absolutely continuous part of Λ_t is denoted by Λ_t^c.

On the set $E_0 = E_X \times (E_A \cup \{\varnothing\}) = \{(x, a) \in E : \ x \neq \varnothing\}$, i.e. when covariates are measured, the local description $d\Lambda_t(dz)$ may be written in a way that clearly reflects the time order in which covariate measurements and actions have taken place,

$$\begin{aligned} d\Lambda_t(dz) &= d\Lambda_t(E_0) \ \frac{d\Lambda_t(dx, E_A \cup \{\varnothing\})}{d\Lambda_t(E_0)} \ \frac{d\Lambda_t(dx, da)}{d\Lambda_t(dx, E_A \cup \{\varnothing\})} \\ &= d\Lambda_t(E_0) \ \phi_t(dx) \, \psi_{t,x}(da) \end{aligned}$$

with $dz = dx \times da$. Here $d\Lambda_t(E_0)$ is the compensator for the underlying counting process $N_t(E_0)$ that jumps every time there is a covariate measurement. If the history $\mathcal{F}_t$ has the form $\mathcal{F}_t = \mathcal{F}_0 \vee \tilde{\mathcal{F}}_t$, the function $\phi_t(\cdot)$ can be interpreted as the conditional distribution of measured covariates at time t, given that there is such a measurement, whereas $\psi_{t,x}(\cdot)$ has the interpretation as the conditional distribution of the action, given the strict pre-t history $\mathcal{F}_{t-}$ and the information that a covariate measurement was done at time t, then giving the result x.

On the set $E \setminus E_0 = \{\varnothing\} \times E_A$, i.e. when covariates are *not* measured but an action is taken, then $d\Lambda_t(dz) = d\Lambda_t(\varnothing, da)$ is rewritten as

$$\begin{aligned} d\Lambda_t(\varnothing, da) &= (1 - \Delta\Lambda_t(E_0)) \ \frac{d\Lambda_t(\varnothing, da)}{1 - \Delta\Lambda_t(E_0)} \\ &\doteq (1 - \Delta\Lambda_t(E_0)) \ d\Lambda_t^0(\varnothing, da) \end{aligned} \tag{2}$$

Here $1 - \Delta\Lambda_t(E_0)$ is the probability of not getting a covariate measurement at time t, which is equal to 1 if the compensator $\Lambda(E_0)$ does not have a jump point at t. This would correspond to a design in which new covariate measurements appear at times which are completely unpredictable on the basis of the histories $(\mathcal{F}_t)$.

3. Causal reasoning and prediction

The philosophical problem of what causality means, and under what circumstances, if any, an observed sequence of events can be given a causal interpretation, has given rise to a vigorous debate which has continued for centuries. Ultimately, the problem might well remain unsettled. An important ingredient in this debate is Hume's (Hume, 1739) legacy that proof is impossible in empirical science, as well as contributions from philosophers

such as Bayes (1764), Popper (1959), Kuhn (1962) and Suppes (1970). A more elaborate discussion of this matter, although interesting, is beyond the scope of this paper however.

Instead, we shall here concentrate on two important aspects of causality. The first aspect is temporality, which refers to the necessity that a cause must precede the effect in time. Although it is difficult to argue against this basic postulate, it is often overlooked in the design and analysis of observational studies.

Secondly, not all study designs allow statistics to play a part in the process of causal reasoning. In some designs one can simply not distinguish whether the observed change in response is due to an action taken or rather due to differences between unobserved attributes of different treatment groups also known as confounding.

The requirement of temporality is implicitly fulfilled in our framework to the extent that the response is ultimately measured at the end of the study period. If the response, however, is a function of the observed data, e.g. total time with CD_4-counts below level c, one has to cautiously investigate the issue of temporality in each particular case. In the above example the response should be defined as the total time *after* possible treatment initiation.

In the following a response can either be a function of the observed data or an unobserved future response. To include both cases in the framework we provide here a general definition of a response.

Definition 1 A random element Y with values in a space $(V, \mathcal{V})$ is called a response if Y is defined on the probability space $(\Omega, \mathcal{F}, P)$.

In the simplest case Y is just a measurement taken on, or the status of, one or more individuals at the end of follow-up, while in a more complicated case, Y might be the number of CD_4-measurements below a certain threshold level c counted over a time interval, or the total time under treatment, or a time varying marker being a part of the covariate process, or even an unobserved future outcome.

A necessity in empirical research, beyond investigating when causal reasoning is possible (see Section 3.1.), is how to quantify the magnitude of the causal effect. This problem has roots in both statistics and philosophy. Here we narrow our discussion to concern the extent in which probabilities can be used in quantifying causal effects.

Investigating a hypothesis like "does A cause Y", one is often in reality asking "is A frequently followed by Y" rather than "is A always followed by Y". For example, "tobacco smoking causes lung cancer" does not imply that once a person starts smoking he/she will some day suffer from lung cancer.

It has become clear that smoking by itself is not a sufficient cause but that other factors are required to be present before lung cancer will manifest itself. A philosophical question, related to the discussion of whether the world is deterministic or stochastic, is whether it is possible to measure a sufficient amount of other related factors such that, controlling for these, tobacco smoking will *always* cause lung cancer.

For the causal effect itself this can be formulated as the question of whether or not there (always) exists an individual causal effect. While some framework for causal analysis relies on such existence, we have avoided it here.

Rather than considering individual causal effects we restrict our attention to the interpretation of probability as a property for an equivalence class of exchangeable individuals. Hence, if we are considering two exchangeable individuals having the same observed history, we would predict the same response if the same actions were taken. This way of interpreting probability is consistent with both "A is frequently followed by Y" and "A is always followed by Y", since in the latter case, if the individual causal effects really existed, we could just measure a sufficient amount of related factors to trivialise the prediction.

In reality, though, the causal field of interest is always rather limited and any causal statement derived from probability will, except from rare cases, be of the type "A is frequently followed by Y".

Prediction, and in particular the work by Arjas and Eerola (1993), becomes a possible ingredient in quantifying causal effects. For the sake of completeness, we shall here provide only a very short summary of predictive distributions.

The basic element in predictive distributions is the *prediction process* $(\pi_t)_{t\geq 0}$ defined by

$$\pi_t(B) = P(Y \in B \mid \mathcal{F}_t), \quad B \in \mathcal{V} \tag{3}$$

and being the probability of later observing the event $\{Y \in B\}$ conditionally on information available at time t. If we denote by $F_t = \{(T_n, Z_n); T_n \leq t\}$ the observed pre-t history, it is possible to show that there exists regular versions of transition probabilities $\pi_t^*(\cdot,\cdot)$ such that $\pi_t(B) = \pi_t^*(F_t, B)$ (i.e. $P(Y \in B \mid \mathcal{F}_t)(\omega) = \pi_t^*(F_t(\omega), B)$). This implies that predictions based on $(\mathcal{F}_t)$ can be calculated from the observed data.

Now predicting the outcome and thus exploring the causal relationship between an action and a response can be assesed through the prediction process in three different ways. First, we might look at the distribution $B \to \pi_t^*(F, B)$ for given history F at time t, which is the simple matter

of predicting Y based on observed data. Second, we could also look at the process $t \to \pi_t^*(F, B)$ for given (F, B), expressing an effect of learning from the data, that is, how the probability of $\{Y \in B\}$ is updated in time based on progressive observation of (F_t). Third, and last, we could explore $F \to \pi_t^*(F, B)$, expressing how the probability of $\{Y \in B\}$ depends on what happened before t, mimicking the idea of counterfactuals.

It is this last way of reasoning which is perhaps most naturally used for drawing causal conclusions from empirical study. We can choose t to be a time when all the available data has been observed, and then consider predictions concerning the future of one or more hypothetical individuals with a pre-assigned past. (A slightly different possibility is to consider an individual actually included in the data, but such that the follow-up has been right-censored, or artificially "removed" from the data in order not to make the prediction tautological.) A particular novelty of Bayesian inference in all such considerations is that there is no separation between statistical estimation and assessement of causal influences: the resulting posterior is integrated directly into computation of the predictive distributions.

All three mappings are important tools when investigating causal effects that normally have a time-dynamic aspect. For example, some treatment may be toxic and therefore considered harmful from a short term perspective but beneficial when expanding the time horizon. For detailed examples of the three mappings we refer to Arjas and Eerola (1993), Eerola (1994), and to Arjas and Andreev (1999) for an application of the predictive distributions.

3.1. Confounding

A fundamental principle in empirical studies directed towards causal conclusions is the principle of controlling for confounding. When assessing the effect of an action on the considered response one should control for factors that might have the potential of confounding the effect of interest. The intuition behind this principle is the perception that an action will have a causal effect on a response if the response changes when the action changes "under circumstances in which everything else remains unchanged". In other words, fixing potential confounders is an attempt to isolate the causal mechanism from possible alternative ways of explaining the response. It is of course not possible to perform such ideal experiments in reality, and in nonexperimental observational studies the ideal is even further away from what can actually be accomplished. Nevertheless, this general principle of controlling for confounding is a focal point in all causal argumentation. Here it gives rise to a natural formulation of when causal reasoning is possible, without having to assume the existence of individual level causal effects or potential outcomes.

Let (U_t) be a process of unobserved variables and denote by $\mathcal{G}_t \subset \mathcal{F}$ the pre-t history generated by (U_t). Let $(\mathcal{F}_t)$, with $\mathcal{F}_t = \tilde{\mathcal{F}}_t \vee \mathcal{G}_t$, be the filtration containing the information of both the observed history of the marked point process and the history of (U_t). To facilitate a likelihood based argument for when causal reasoning is possible, we shall henceforth assume that (U_t) is left continuous and hence predictable from $(\mathcal{F}_t)$.

By saying that (U_t) consists of unobserved variables we think of both the case where (U_t) is "potentially observable", like covariate measurements that could in principle be collected but were not, and the case where the U_t's are genuinely not observable. Thus (U_t) could represent individual frailties, or even "structural" model parameters, say θ, which in classical statistical inference are commonly interpreted as characteristics of a population. At this point we do not make a distinction between unknown parameters and unobserved random variables, apart from the convention that the former are supposed to be fixed in time and therefore measurable already with respect to $\mathcal{G}_0$. Should one want to consider a statistical model $\{P^\theta;\, \theta \in \Theta\}$ parametrized according to such θ, it can always be realized in terms of regular conditional probabilities arising from P and with the value of θ given.

Resembling the definition of confounding by observed variables we provide here the definition of *no unmeasured confounders* in the present framework. This definition bears a close correspondence, both technically and conceptually, to the definition of non-innovative (or non-informative) censoring given in Arjas and Haara (1984), see also Andersen *et al.* (1993).

Definition 2 The action process (A_t) is said to be *not confounded* by (U_t) with respect to $(P, \mathcal{F}_t)$ if at least one of the following two properties holds P-a.s.

(i) The processes $(\psi_{t,x}(\cdot), x \neq \varnothing)_{t\geq 0}$ and $(\Lambda^0_t(\varnothing, \cdot))_{t\geq 0}$ describing the assignment of actions are not utilizing the latent information in $(\mathcal{G}_t)$, in the sense that they remain unchanged if the filtration $(\mathcal{F}_t)$ is changed into the filtration $(\tilde{\mathcal{F}}_t)$ generated by the observations.

(ii) For all $t \geq 0$, Y is conditionally independent of $\mathcal{G}_t$ given $\tilde{\mathcal{F}}_t$.

Of these two criteria (ii) has a more obvious intuitive justification: Consider a situation at "the present" time t at which we try to predict the value of the future response Y. The process (U_t) should obviously not be called a confounder in the causal problem at hand if knowing its entire development up to the present would not change the prediction concerning the future response Y beyond what can be said on the basis of the observations actually

made. Thus, even in situations in which (U_t) is actually observed and therefore its development up to the present time is known, we would normally not include it in a causal model for Y if it satisfies (ii).

Postulate (i) looks more technical and the intuition leading to this formulation is also more subtle. The distributions $\psi_{t,x}(\cdot), x \neq \varnothing$, and $\Lambda_t^0(\varnothing, \cdot)$ representing local assignment of actions over $E_A \cup \{\varnothing\}$ satisfy (i) if and only if they do not depend on $\{U_s,\ s < t\}$. At time t the past observed covariate values and past actions are known, and thereby fixed. Suppose that of the covariate-action pair (x, a) the value of x is always checked first by observation, after which the action a is determined. As before, we count $(\varnothing, a)$ with $a \neq \varnothing$ ("action without preceding covariate measurement") as a marked point, but not $(\varnothing, \varnothing)$. Different actions, including $a = \varnothing$ ("no action is taken"), may in general lead to different values of the response Y, and this is reflected in the predictive distribution of Y that can be issued at time t. It will be justified to call such dependence on the action taken, or not taken, causal if the effect is not mixed (confounded) by some unobserved factors already present at or before time t and also contributing to the value of the response Y. While causal influence is only possible forwards in time, the unobserved past $\{U_s,\ s < t\}$ could in principle very well depend probabilistically, in the sense of "inverse probability", on an action taken at time t. Postulate (i) rules out such dependence, both when some $x \neq \varnothing$ is observed at time t and when there is no such x.

It should be noted that the likelihood justification of Definition 2 given in Section 4. below allows a particular design to alternate between requirements (i) and (ii) over time in such a way that at any given time at least one of them holds.

Since (U_t) can be chosen arbitrarily we say that the design has *no unmeasured confounders* if the conditional independence postulate of Definition 2 holds for every candidate for (U_t), that is, for every potential confounder process. Controlling a potential confounder process by observation, whenever possible, is of course a direct way in which one can try to assure that the design has no unmeasured confounders.

In practice, suggesting candidates for (U_t) is limited to the *causal field* under consideration, and therefore to unobserved processes that the investigator is conjecturing, or knows, to have an effect on Y. Any statement about causal reasoning would then be relative to the set of variables constituting the considered causal field. In practice, however, one must limit both the number of different covariates being measured and the amount of unobserved variables included in the causal field in order to justify the "no unmeasured confounders" assumption.

By measuring more and more covariates the validity of the conditional

independence assumptions of Definition 2 will naturally become more likely, while at the same time, including more unobserved variables into the causal field will make them less likely.

It is also important to notice that the assumption of no unmeasured confounders is a conditional independence property between an action and the *past* history of (U_t), and not of the whole process. As a consequence, given the observed past, the choice of a new action does not contribute to the knowledge (inference) concerning the past of (U_t). It is perfectly legitimate that an action taken will (causally) influence the later development of some unobserved variables represented by (U_t), as well as, of course, the later development of observed covariates and, ultimately, the response. Thus we can say that in statistical problems dealing with causality, the influence of actions can be causal but not inferential.

Note also that for covariates we have not assumed such asymmetry; a new covariate reading can have both inferential value backwards in time and causal influence forwards in time. For observed potential confounders this asymmetry in time regarding the action taken, where action depends on previous development in the potential confounders and at the same time predicts their further development, is called *time-dependent confounding* by Robins (1989), Robins *et al.* (1992).

Using the process of unobserved variables we now formulate when causal reasoning is possible using the controlling for confounders principle.

Principle 1 Causal reasoning from data by using statistical methods is possible if and only if there are no unmeasured confounders.

It should be emphasized at this point that the controlling for confounding principle does not imply controlling for every potential confounder of the action process (A_t), but only those that are confounding the effect of interest. Hence, even though a potential confounder has been observed by measurement it does not automatically imply that it should enter the covariate process (X_t). This can be illustrated by two different aspects of conditioning. The first aspect is found in the phase of model building where the investigator has the choice of which covariates to include and thus which to take into account in the conditioning. An example of such choice is whether or not to include *age* in an analysis of the effect of stroke on mortality.

For the second aspect consider a study with several covariates where each considered separately is confounding the effect of interest. When including some of these, but not all, the remaining covariates may satisfy (ii) in Definition 2 and are consequently optional to include.

3.2. Examples

Below we shall give some general examples focusing on different aspects in the assessment of the effect of an action taken.

3.2.1. Discrete time

In a sequential study design, or in discrete time, individual level observations reduce to $(X_{i,1}, A_{i,1}), \ldots, (X_{i,n_i}, A_{i,n_i})$ where X_{ij} or A_{ij}, but not both, can be $\varnothing$. Let U_j be the unobserved potential confounders at time j for all individuals in the sample. In the example above concerning the diagnosis of asthma, peak flow measurements are taken daily for a two week period followed by a physician's decision on whether or not to initiate an asthma treatment on the particular individual. This gives rise to a sequence $(X_{ij}, \varnothing)$, $1 \leq j < 14$, with X_{ij} being the FEV_1 measurement on day j, followed by the treatment decision $(\varnothing, A_{i,14})$ on day 14 .

In discrete time $d\Lambda^0$ in (2) can only be non-zero at time points $j \in \mathbb{N}$ and $\Delta\Lambda_j^0(\varnothing, \cdot)$ then becomes the conditional distribution of action given the strict past and the information that no covariate measurement, i.e. $X_{ij} = \varnothing$, was made. Part (i) in Definition 2 then reads

$$A_{ij} \perp\!\!\!\perp (U_1, \ldots, U_j) \mid (X_{ik}, A_{ik}), 1 \leq k < j, X_{ij}$$

saying that current action should be independent of the past and present unobserved variables conditional on past observed variables and current covariates, if measured, and otherwise on the information that no measurement was made.

3.2.2. Selective treatment

In this example we consider a study of the effect of a new chemotherapy treatment on a number of n cancer patients where the treatment design is such that only a quarter of the patients with lowest immune system status, measured by their CD_4 count, are treated with the new drug. The remaining patients receive a treatment based on an existing drug. Status of the immune system and subsequential reassessment of who to treat, is done weekly. In this sequential treatment study let X_{ij} denote the CD_4 count at week j on individual i and A_{ij} the action taken the same week with value 0 for existing treatment and 1 for new treatment. The response Y of interest is time to remission, also measured in weeks.

Let $(U_t)_{t \in \mathbb{N}}$ be a process of unobserved potential confounders. Since the weekly CD_4 count $(X_{ij}), i = 1, \ldots, n$ on the whole sample determine who will

receive the new treatment, that is (A_{ij}), we have

$$A_{ij} \perp\!\!\!\perp (U_1, \ldots, U_j) \mid (X_{1j}, \ldots, X_{nj})$$

and hence the design has no unmeasured confounders. By controlling for (X_{ij}) the treatment can, at least in principle, be estimated from the data. Note that in this example the assumption of no interference between units does not hold.

3.2.3. Confounded action

The following example, concerning a study of the effect of replacement oestrogens on overall mortality for post-menopausal women, is used to illustrate both the issue of confounding and of predictive distributions.

We consider a number of n post-menopausal women who, at time 0, are randomised to receive either replacement oestrogens or placebo. During follow-up blood cholesterol level is monitored in the treated group by means of scheduled annual measurements performed by their own physician. If a woman undergoing treatment for some other reason visits her doctor, the doctor is instructed to take a blood sample in order to determine and record the current cholesterol level. On the control group no such measurements are made.

One known side effect in the treatment with replacement oestrogens is a possible elevation of the blood cholesterol level. Since elevated cholesterol level is known to be associated with (and perhaps also causally related to) an increase in cardiac mortality, the physicians were instructed to remove these women permanently from the study if elevation was encountered.

Let the treatment process (A_t) take the constant value 1 if a woman is on replacement oestrogens, and value 0 otherwise. The covariate process (X_t) describing the status of the blood cholesterol levels is assigned value 1 if the level is elevated, and 0 if normal. Finally, let the response Y be the age at death.

In this study the intention to treat analysis, comparing treatment as randomised, is unconfounded due to the randomisation, and the intention to treat effect (within the trial) is therefore estimable from data. This effect can be expressed in terms of predictive distributions as the difference between $P(Y \mid \text{protocol treated})$ and $P(Y \mid \text{protocol placebo})$, where the protocol for women assigned to treatment involves measurements of blood cholesterol level and a possible withdrawal of the treatment.

When trying to assess the as treated effect, that is the influence of actions in $P(Y \mid A_1, ..., A_m)$, one should control for any covariate that acts

as a confounder. In this particular study treatment assignment is given as $A_k = 1 - X_k$ among women randomised to replacement oestrogens and hence the effect of action after randomisation is completely confounded by blood cholesterol level. This makes the as treated effect non-separable from the effect of elevated cholesterol level. A comparison of the counterfactuals $Y_{(0,\ldots,0)}$ and $Y_{(1,\ldots,1)}$ (see Robins (1989)) representing the response if never treated and always treated, respectively, neglects the problems of complete confounding and suggests an estimate of the as treated effect that is invalid.

4. Causal reasoning and the likelihood function

As noted already above, condition (ii) in Definition 2 is by itself sufficient to rule out the possibility that (U_t) would be a confounder in the considered causal problem. It is therefore sufficient to consider more formally the implications of condition (i).

The well-known general inferential principles of weak conditionality and weak sufficiency imply that the likelihood function contains all evidence given by the data. It is thus important to justify the connection between Definition 2 and Principle 1 in the important situation in which statistical inference obeys the likelihood principle.

Following the results of Jacod (1975) and using product-integral notation (see Gill and Johansen, 1990) we may write the conditional likelihood of the marked point process on the time interval $[0, \tau]$ given $\mathcal{F}_0$ as

$$P \,|_{\mathcal{F}_0} \prod_{t\in[0,\tau]} \left(\prod_{z\in E} d\Lambda_t(dz)^{\,dN_t(dz)} \left(1 - d\Lambda_t(E)\right)^{1-dN_t(E)} \right) \tag{4}$$

Since $N_t(E)$ makes jumps (of size one) only at times (T_k), (4) can be written as

$$P \,|_{\mathcal{F}_0} \prod_{k\geq 1} \left(\prod_{t\in(T_{k-1},T_k\wedge\tau)} \left(1 - d\Lambda_t(E)\right) d\Lambda_{T_k}(dZ_k)\, 1_{(T_k\leq\tau)} \right) \tag{5}$$

where the product integral $\prod (1 - d\Lambda_t(E))$ can be viewed as the likelihood contribution from "no new marked points on the interval (T_{k-1}, T_k)", and $d\Lambda_{T_k}(Z_k)$ is the likelihood contribution of (T_k, Z_k). Using the representation

$$\prod (1 - d\Lambda) = \exp(-\Lambda^c) \prod (1 - \Delta\Lambda), \tag{6}$$

the product integral in (5) can be rewritten as

$$\begin{aligned}
&\prod_t (1 - d\Lambda_t(E)) \\
&= \prod_t (1 - \Delta\Lambda_t(E_0)) \, \exp(-\Lambda^c_{T_k \wedge \tau}(E_0) + \Lambda^c_{T_{k-1} \wedge \tau}(E_0)) \\
&\quad \prod_t (1 - \Delta\Lambda^0_t(E \setminus E_0)) \, \exp(-\Lambda^c_{T_k \wedge \tau}(E \setminus E_0) + \Lambda^c_{T_{k-1} \wedge \tau}(E \setminus E_0)) \\
&= \prod_t (1 - d\Lambda_t(E_0)) \, (1 - \Delta\Lambda^0_t(E \setminus E_0)) \\
&\quad \exp(-\Lambda^c_{T_k \wedge \tau}(E \setminus E_0) + \Lambda^c_{T_{k-1} \wedge \tau}(E \setminus E_0)) \qquad (7)
\end{aligned}$$

where all product integrals are over $(T_{k-1}, T_k \wedge \tau)$. Now, due to (1), the equality

$$\exp(-\Lambda^c_t(E \setminus E_0)) = \exp(-\Lambda^{0,c}_t(E \setminus E_0))$$

holds, and then (i) in Definition 2 implies that all but the first term $\prod(1 - d\Lambda_t(E_0))$ on the right hand side of (7) remain unchanged when the filtration is changed from $(\mathcal{F}_t)$ to $(\tilde{\mathcal{F}}_t)$. Thus, in likelihood based inference concerning the past of the unobserved process (U_t), these terms can be treated as constants not depending on the parameter of interest. Note that this coincides with the likelihood contribution which we would get from this interval if we only paid attention to the new covariates, that is, in our inference we can replace "no new points in E" by "no new points in E_0".

Consider then the differential term in (5). In the case where $X_k \neq \varnothing$, we have

$$d\Lambda_{T_k}(dZ_k) = d\Lambda_{T_k} \, \phi_{T_k}(dX_k) \, \psi_{T_k, X_k}(dA_k),$$

where the first part of (i) in Definition 2 implies that $\psi_{T_k, X_k}(dA_k)$ can be treated as a constant proportionality factor, leaving only the contribution of the "marginal" $(X_k, \cdot) \in E_0$. If, on the other hand, $X_k = \varnothing$, we must have $A_k \neq \varnothing$, in which case we write the k'th factor in the likelihood as

$$\begin{aligned}
d\Lambda_{T_k}(\varnothing, dA_k) &= (1 - \Delta\Lambda_{T_k}(E_0)) \, \frac{d\Lambda_{T_k}(\varnothing, dA_k)}{1 - \Delta\Lambda_{T_k}(E_0)} \\
&\propto (1 - \Delta\Lambda_{T_k}(E_0))
\end{aligned}$$

where the proportionality again follows from part (i) of Definition 2.

In summary, up to proportionality in the potential confounder variables, we have therefore established

$$\prod_{t \in (T_{k-1}, T_k \wedge \tau)} (1 - d\Lambda_t(E)) \, d\Lambda_{T_k}(dZ_k)$$

$$\propto \begin{cases} \prod_{t\in(T_{k-1},T_k\wedge\tau)}(1-d\Lambda_t(E_0))\, d\Lambda_{T_k}(dX_k,\cdot) & \text{on } \{X_k \neq \varnothing\} \\ \prod_{t\in(T_{k-1},T_k\wedge\tau]}(1-d\Lambda_t(E_0)) & \text{on } \{X_k = \varnothing\} \end{cases}$$

Note here that, if $X_k = \varnothing$, the contribution of the marked point $(\varnothing, A_k)$ disappears completely.

Therefore, in the inferential problem concerning the values of the potential confounder in the past, we can on each interval $(T_{k-1}, T_k]$ ignore any possible role of actions, that is, both an action A_k actually taken, as well as the fact that no actions were taken on (T_{k-1}, T_k).

In a typical situation, the σ-algebra $\mathcal{F}_0 = \mathcal{G}_0$ contains certain random variables whose values are supposed to be fixed at time zero and thus constant during the study. Examples of such variables are the randomisation indicator in a randomised clinical trial, and structural parameters in general. Structural parameters θ can in frequentist inference be thought of as being a realization of a random variable ζ describing e.g. population properties. In this case, as already noted above, the probability measure P^θ is given as $P^\theta(\cdot) = P(\cdot \mid \zeta = \theta)$ and $\sigma(\zeta) \subset \mathcal{F}_0$.

In Bayesian inference structural parameters are treated as random variables and viewed as a part of the underlying state of nature, and would therefore be described as elements in the (U_t)-process.

4.1. No interference between units

In the above discussion of causal reasoning, we did not make it explicit whether we were considering just a single individual or a sample of individuals upon which inference would then be based. If we are considering just a single individual, the history $\tilde{\mathcal{F}}_t$ represents knowledge regarding this particular individual whereas $\mathcal{G}_t$ may be linked to background population variables.

If, on the other hand, we are following m different individuals, the internal history $\tilde{\mathcal{F}}_t$ of the marked point process may have a more complex structure since the decision of what action to take on a specific individual may also depend on knowledge from treating other individuals in the sample. An example is a trial, say two-armed testing of a new treatment against an existing one, where it would be considered unethical to offer a patient anything but the best treatment judged on the basis of his/her current status. Over time, the physician will gain knowledge from other patients on the efficacy and side effects of the new treatment. The choice of how to treat a certain patient will then depend on the history of all the other patients. Assuming independence between individuals would in this case be too strong an assumption.

The idea of *no interference between units* (Cox, 1958) or *SUTVA* (Rosenbaum and Rubin, 1983) has an analogue in continuous time, which can be stated as

$$Y_i \perp\!\!\!\perp A_{jl} \mid \{(T_{ik}, Z_{ik}); 1 \leq k \leq n\}, \ \ \forall n, \forall j \neq i, \forall l : T_{jl} \leq T_{in}$$

for $A_{jl} \neq \varnothing$. The interpretation is that, for given covariate measurements, the influence of an action on the response for individual i does not depend on what actions were taken on other individuals. This is for instance not true in the case of a chicken pox epidemic in a community when parents are considering whether to take their child out from municipal day care into home care. The effect of such an action on infection free time will then depend on what the other parents decide to do.

Although the assumption of no interference between units is not necessary to allow causal reasoning, it does simplify the statistical inference a great deal.

5. Connections to other causal models in the literature

Several alternative frameworks for causal reasoning have been suggested earlier in the statistical literature. The following is an attempt to see how they connect with the framework presented here.

5.1. The Rubin causal model

In the Rubin causal model only one action is involved, allowing it to be formulated in the present framework as follows. For a given individual, covariates X are measured at time 0 and this is immediately followed by an assignment A to either treatment or placebo. At a later point in time τ, a response Y is measured.

Following the idea of the Rubin causal model, there exist for each individual i and already before the experiment takes place, two potential outcomes, namely the potential outcome $Y_0(i)$ under placebo and the potential outcome $Y_1(i)$ under treatment. The assignment indicator determines which potential outcome will be observed. Furthermore, the *individual causal effect* for individual i is defined as the non-random quantity $Y_1(i) - Y_0(i)$.

The randomness in the model is assumed to be caused by heterogeneity among individuals, giving rise to population level potential outcomes Y_0 and Y_1 with respective distributions given as $Y_a \stackrel{\mathcal{D}}{=} (Y \mid A = a)$, $a = 0, 1$. By assuming that the treatment assignment is strongly ignorable, that is,

$(Y_0, Y_1) \perp\!\!\!\perp A \mid X$, the average (over population) causal effect can be estimated without bias.

The problem with connecting the Rubin causal model to the framework presented in this paper is twofold. Firstly, the existence of the individual potential outcomes before the experiment takes place is crucial to his model, and secondly, randomness is only due to heterogeneity among individuals.

Since the present framework is solely based on random variables and their distributions, it does not provide a way of formulating the idea of individual potential outcomes and hence there is no direct way of comparing the ignorable treatment assumption with our assumption of no unmeasured confounders. Instead we shall make the connection in an intuitive way, using the common interpretation of the ignorable treatment assumption as *the assignment A not "carrying" any information about the future outcome Y other than what is already measured by the covariates X*. Discarding the a priori given individual potential outcomes and just looking at their distribution over a population we can connect the Rubin causal model to our framework by reducing the process (U_t) of unobserved variables to a single variable U (equal to U_0) and then consider the two different situations, one being U influencing Y directly and the other U influencing Y only through X and A.

In the first case, we could set $U = (Y_0, Y_1)$ and interpret the population level potential outcomes, rather than arising from the individual potential outcomes, as unobserved variables influencing the response directly and not only through measured covariates and the assignment. For this particular choice of U the ignorable treatment assumption is obviously the same as (i) in our Definition 2. The second case, where unobserved variables are influencing the response only through measured covariates and the assignment ((ii) in Definition 2), is consistent with the interpretation of ignorable treatment assignment but does not allow an explicit formulation in terms of the (Y_0, Y_1). So, the intuitive idea of when causal reasoning is possible using statistical methods is the same in the two models.

As mentioned earlier, randomness in the Rubin causal model is assumed to be caused only by heterogeneity among individuals. This is different from the framework presented here where other sources of randomness, like measurement error and subjective uncertainty, are plausible.

5.2. The Rubin causal model in sequential trials

An important special case of the present framework is the sequential trial design studied by Robins (1986, 1997), Robins *et al.* (1992). It is a sequential version of the Rubin causal model, where individuals have their covariates X_l observed at predetermined time points $l = 0, 1, \ldots, K$ imme-

diately followed by an action A_l. By the end of the follow-up, at time K, a response Y is measured. Hence, data on a single individual i consist of $\{(X_{i0}, A_{i0}), \ldots, (X_{i,K_i-1}, A_{i,K_i-1}), Y_i\}$ with different such items assumed to be drawn independently from a common distribution. For a given treatment regime $a = (a_0, \ldots, a_K)$, Y_{ia} represents the individual potential outcome of Y_i when treatment a is applied and Y_a has the the distribution of Y_{ia} over the population.

As in the Rubin causal model, the individual causal effect is assumed to exist and Y_{ia} to be defined a priori, before the experiment. To facilitate unbiased estimation of the average causal effect, the assumption of ignorable treatment (also called "no unmeasured confounders"),

$$Y_a \perp\!\!\!\perp A_l \mid (X_0, \ldots, X_l, A_0, \ldots, A_{l-1})$$

whenever $P(A_l = a_l \mid X_0, \ldots, X_l, A_0, \ldots, A_{l-1}) > 0$, is imposed.

Having made the intuitive connection between ignorable treatment assumption and Definition 2 in the case where $K = 1$ (the Rubin causal model), it follows immediately that the ignorable treatment assumption in the sequential trial design, which is a special case of the design considered in this paper, has the same intuitive justification as our no unmeasured confounders hypothesis.

One aspect that is not covered by the above sequential trial design but is feasible using the marked point process approach is to take into account the times *when* in time individual measurements are sampled.

5.3. Potential outcomes and the G-computation formulae

The philosophical issue (see Sosa and Tooley, 1993) of whether the terminology of potential outcomes or counterfactuals is necessary for any discussion of causation has in the past given rise to a very controversial debate. It is clear that counterfactuals are an appealing way to quantify an effect that has a possible causal interpretation. This, however, is a separate issue from when causal reasoning by applying statistical inferential tools would be possible.

We shall in this subsection proceed with more concrete examples of how to express a causal effect of an action taken. As we shall see, all these examples are handled naturally by applying the idea of predictive distributions. As a matter of fact, statistical methods can never provide anything more than what can be expressed by predictive distributions.

The hypothetical comparison of the actual observed outcome with what would have happened had we given the patient another treatment is tempting when dealing with a study design such as the one in the Rubin causal model. Except for a changed action, the rest of the experiment remains the same.

From a probabilistic point of view we do not need to assume the existence of a unit causal effect, neither a priori existence of potential outcomes. Due to lack of identifiability on an individual level, the comparison of the observed outcomes with their counterfactual counterparts on a population level is just a matter of investigating the change in marginal distribution of the outcome when changing the assignment, i.e. the change from $P(Y_{a_1}) = P(Y \mid A = a_1)$ to $P(Y_{a_2}) = P(Y \mid A = a_2)$ when $a_1 \neq a_2$.

When dealing with repeated actions on a single individual over time this comparison of counterfactuals, and hence the marginalised outcome distributions, may turn out to be too rigid to provide a useful answer in an attempt to estimate the effect of an action. For some questions, as in the next example, the study design has a dynamic structure not captured by the comparison of static treatment regimes expressed in terms of counterfactuals.

Consider the problem of comparing two different types of cancer treatments. Due to the differences among patients and their tumours, having a non-dynamic predetermined future treatment as considered by the counterfactuals, Y_a is not providing information on what treatment the physician should offer to the patient. More realistically, one would like to compare two protocols on how to treat patients, where the actual treatment always depends on the current status of the patient. This cannot be expressed through the counterfactuals, but is easily formulated using the present framework and by comparing $P(Y \mid \text{protocol 1})$ with $P(Y \mid \text{protocol 2})$, both being estimable from the data.

Another aspect of prediction is the following. Given current information on a patient and on his/her former disease and treatment history, the physician will naturally want to base the choice of a future treatment on what he/she is predicting to be the outcome. At time T_k, this corresponds to comparing $P(Y \mid \{(T_i, Z_i); 1 \leq i \leq k-1\}, T_k, X_k, (A_k = a))$ for different choices of action a. We might think of the properties of A_k implied by no unmeasured confounders as the action A_k being freely manipulable without knowledge about the U's up to time T_k, allowing us to perform the operation of *setting* the value of action without changing the circumstances of the study (see Pearl, 1995).

An important issue when using counterfactuals is how to calculate their distribution from observed data. Assuming ignorable treatment in a sequential trial, Robins (1986, 1989, 1998) introduces a procedure called the G-computation algorithm for calculating the distribution $P(Y_a)$ of the counterfactual as $P(Y \mid A_0 = a_0, \ldots, A_{K-1} = a_{K-1})$, carried out by integrating over the distribution of covariates. This requires explicit modelling of the conditional distributions $P(Y \mid (X_i, A_i)_{\,0 \leq i \leq K-1})$ and $P(X_{n+1} \mid (X_i, A_i)_{\,0 \leq i \leq n})$ for $0 \leq n \leq K-1$, but clearly not $P(A_{n+1} \mid (X_i, A_i)_{\,0 \leq i \leq n}, X_{n+1})$. Observe that

fixing the treatment regime to a is the same as fixing the treatment part of the history $(\mathcal{F}_t)$ in (3), and calculating the distribution of the counterfactual Y_a amounts to calculating the predictive distribution $\pi_t(B)$ for a given time $t = T_K$ and history $(A_0, A_1, \ldots, A_{K-1})$.

The G-computation algorithm is really nothing more than using the chain rule for conditional distributions on $P(Y \mid A_0 = a_0, \ldots, A_{K-1} = a_{K-1})$, where the connection between the counterfactual and the observed outcome is made by successive use of the no unmeasured confounders assumption. Hence, following the axioms of probability theory, the G-computation algorithm is not only an obvious choice but in fact the only one that is consistent. A procedure similar to the G-computation algorithm is also found in Pearl (1995). In the marked point process framework, with even more complicated designs than the sequential trial, the distribution of counterfactuals is of course calculated in a similar fashion.

Strictly speaking, the G-computation algorithm in the above mentioned references is only formulated for observed covariates and therefore does not cover models having unobserved frailty parameters. Predictive distributions do not have such a restriction.

5.4. Graphical models

Graphical models provide a convenient tool for exploring and for the visualisation of dependence relations between random variables. This is especially true for conditional independencies since they can be read off directly from the graph. Although directed acyclic graphs often reflect the time order in which sampling on a single individual is performed, they do not fully incorporate the time aspect. More specifically, they do not specify *when* in time measurements and actions were taken. This is crucial if inference is based on a sample of individuals and sampling is not following the sequential trial design with predetermined time points.

It is, however, still possible to use the directed acyclic graphs to illustrate the idea of the no unmeasured confounders assumption if we restrict the model to a single individual and graph the data *after* sampling. As before, we have measured data $(T_i, Z_i)_{1 \le i \le n}$ and are considering a response Y that is either observed or defined upon the observed data. Let $(U_1, \ldots, U_n)$ be a vector of unobserved variables possibly confounding the effect of actions on the response, where $U_k = U_{T_k}$ is the vector of unobserved variables at time T_k. From the nodes $(U_k, X_k, A_k); 1 \le k \le n$, one can readily create a corresponding saturated directed acyclic graph with arrows going only forwards in time. In this graph, the assumption of no unmeasured confounders is the same as saying that at every time point T_k either the edge $U_k \to A_k$ or the

edge $U_k \to Y$ can be removed without changing the joint distribution defined on the graph.

Without discussing the definition of causal effects given in Pearl (1995), the no unmeasured confounders assumption can be seen to be equivalent to Pearl's identifiability of causal effects by the following argument. At every time point k, conditionally on the history up to $k-1$, the graph above is the same as the union of the graphs (c) and (d) in Pearl's Figure 6. Removing $U_k \to A_k$ implies graph (c), while the removal of $U_k \to Y$ implies graph (d). Since the graphs in Figure 6 are maximal, in the sense that the introduction of any additional arc or arrow onto an existing pair of nodes would render non-identifiability of the causal effect as exemplified by Figure 7 (d)+(e), we get the equivalence between Pearl's identifiability of causal effects and our no unmeasured confounders.

5.5. Granger causality

The issue of the role of statistical models and methods in causal reasoning is not only restricted to biometry and epidemiology but has indeed spawned a rich literature in the field of econometrics where special focus has been on the ideas of Granger (1969) and Sims (1972). While the central question in biometry most often is whether (A_t) causes Y, with Y being a fixed target, it is the question of noncausality in econometrics and specifically whether it is possible to reduce the available information in order to predict the future development of a given stochastic process, say (Y_t).

Until recently, noncausality was only discussed for time-series indexed by a discrete time set with the main focus on forecasting in mean, either one-step ahead (Granger's type definition) or for any horizon (Sims' type definition). For equivalence between Granger and Sims causality we refer to the work by Chamberlain (1982) and Florens and Mouchart (1982, 1985).

Here we shall briefly discuss the difference between Granger causality, as it was proposed originally, and the framework presented here, and then make the connections to a later proposed continuous time approach which in a special case is similar to ours. It should be emphasised that Granger causality does not contain the element of action or manipulation, which is why our discussion below does not involve the corresponding process (A_t). Instead, we will consider the response Y as a process (Y_t) which is progressively observable in time.

To fix ideas consider the discrete time processes $(U_t)_{t\in\mathbb{N}}$, $(X_t)_{t\in\mathbb{N}}$, $(Y_t)_{t\in\mathbb{N}}$. As in Engle *et al.* (1983), we do not restrict ourselves to time series or linear models. Granger one-step ahead noncausality can in this setup be expressed

as

$$Y_{t+1} \perp\!\!\!\perp (U_1, \ldots, U_t) \mid (X_1, \ldots, X_t, Y_1, \ldots, Y_t) \tag{8}$$

for all $t \in \mathbb{N}$, saying that (U_t) is not causing (Y_t). Important for the concept of causality in the Granger sense is not the issue of confounding of effects but rather whether knowing the past of the process (U_t) would change the prediction of (Y_t) one time unit ahead. Thus the conditional independence (8) can be thought of as expressing a property of predictive sufficiency of the data. This shows that Granger causality is linked closely with the requirement (ii) of our Definition 2.

To make this connection even clearer, consider the notion of strong global noncausality in continuous time as formulated in Florens and Fougere (1996). (For earlier formulations of these same ideas, see Schweder (1970), Mykland (1986) and Aalen (1987)). They consider a specific process (Y_t) generating $(\mathcal{Z}_t)$, the smallest of three filtrations of increasing size, the others being our $(\tilde{\mathcal{F}}_t)$ generated by the observations including the development of (Y_t), and $(\mathcal{F}_t)$ with $\mathcal{F}_t = \tilde{\mathcal{F}}_t \vee \mathcal{G}_t$ where $(\mathcal{G}_t)$ is generated by the unobserved process (U_t). The filtration $(\mathcal{F}_t)$, and thus the unobserved process (U_t), is said to not strongly cause (Y_t) given $(\tilde{\mathcal{F}}_t)$ if the conditional independence

$$\mathcal{Z}_{t+h} \perp\!\!\!\perp \mathcal{G}_t \mid \tilde{\mathcal{F}}_t$$

holds for all $t \geq 0$ and all positive horizons h. This is almost the same as Definition 2 (ii), and actually identical to it if we take the response Y to be something whose values can be resolved on the basis of histories of (Y_t) of a finite length.

6. Discussion

The main issue in this paper has been our attempt to clarify when and how statistical inferential tools can play a role in establishing causal conclusions. In this process it is important to separate these questions from deeper philosophical problems such as the existence of a unit causal effect. Whatever the answer might be, it does not fundamentally influence the way in which causal conclusions can be drawn from empirical study.

It is inarguable that the human interpretive element in causality is of great importance, and, as pointed out already by Hume (1748), that comparison of counterfactual outcomes is to the human mind a natural way in which the concept of causality manifests itself. But both the observed and the counterfactual outcome cannot exist in the real world, and causality, therefore, always necessitates a mental construction of some form. It is questionable

whether counterfactuals form the natural basis for the methodology of causal reasoning in statistics, however. From a probabilistic point of view one can easily create a probability space large enough to have both the observed and the individual counterfactual outcome defined, thus allowing the comparison of counterfactual outcomes. But again, on the level of real individuals, this difference is unidentifiable from observed data regardless of the methods which one might want to apply. Thus, from the point of view of application, if causal effects are to be quantified by comparing the distribution of counterfactuals, then using a framework based on a priori existing individual causal effects defined upon counterfactual outcomes, as in the Rubin causal model, does not as such imply a change in statistical analysis, nor that different causal statements would be made on the basis of such analysis.

It is generally accepted that the randomisation device, at least in principle, will lead to valid causal inferences, whereas results from observational studies normally are given the lower status of statistical associations. Even then, such inferences are in practice based on an observed finite sample and will therefore depend on how well the realization of the randomisation managed to distribute the unobserved characteristics over treatment groups. In small samples this can be a problem. Thus, while the condition of Definition 2 is guaranteed to hold in designs in which treatment assignment is through a genuine randomisation device, this will only protect us against biasing our causal conclusions systematically by unwarranted confounding. In concrete applications, randomisation is therefore a valid argument for claiming that the analysis is not confounded, but in no way canonical nor exclusive. In a particular observational study one can have high confidence in the analysis not being confounded and hence allow valid causal conclusions to be drawn.

A problem often encountered in observational studies is the issue of temporality. By definition, a cause has to precede the effect. This imposes great practical difficulties into how to deal with causality particularly in cross-sectional studies. In such studies values of variables confounding the action are often measured at a later time point than the action taken, and then the choice of the action can have influenced the values that were observed. Here, instead of correcting for confounders, one is likely to be correcting for intermediate variables on the causal path from action to response, in conflict with another basic principle in empirical research.

Finally, it was argued that all statistical analysis of causality, in the sense of effects of causes (cf. Dawid, 1995), boils down to considering and comparing predictive distributions. As a consequence, the content of causal statements will depend on how probability is interpreted in the context. In the frequentist mode of inference, probability is typically taken to describe the randomness when sampling individuals from a heterogeneous population, or

the variation in hypothetical repeated experiments or trials on such individuals. In Bayesian inference, predictive probabilities are essentially a quantification of the subjective uncertainty regarding unobserved characteristics of members of an equivalence class consisting of individuals, which, based on their observed characteristics, are considered exchangeable. Thus, whatever interpretation of probability is adopted, the causal analysis does not necessitate the assumed existence of individual causal effects.

Acknowledgements

This research was conducted at the Rolf Nevanlinna Institute, University of Helsinki, while the first author held a travelling grant from the Nordic Network for Biostatistics Research. This study was furthermore supported by contract 9602067 from The Danish Research Councils and partly by contract 2R01 CA54706-04A1 from the National Cancer Institute.

References

Aalen, O. O. (1987) Dynamic Modelling and Causality. *Scand. Act. J.*, 177–190.

Andersen, P. K., Borgan, Ø., Gill, R. D., and Keiding, N. (1993) *Statistical Models Based on Counting Processes.* New York: Springer.

Arjas, E. (1989) Survival Models and Martingale Dynamics (with discussion). *Scand. J. Statist.*, **16**, 177–225.

Arjas, E., and Andreev, A. (1999) Predictive inference, causal reasoning, and model assessment in nonparametric Bayesian analysis: a case study. *Submitted.*

Arjas, E., and Eerola, M. (1993) On predictive causality in longitudinal studies. *J. Statist. Plan. Inf.*, **34**, 361–386.

Arjas, E., and Haara, P. (1984) A Marked Point Process Approach to Censored Failure Data with Complicated Covariates. *Scand. J. Statist.*, **11**, 193–209.

Bayes, T. (1764) An essay towards solving a problem in the doctrine of chances. *Phil. Trans. R. Soc.*, **53**, 370–418 (reprinted in *Biometrika* (1958), **45**, 293–315).

Brémaud, P. (1981) *Point Processes and Queues.* New York: Springer.

Chamberlain, G. (1982) The General Equivalence of Granger and Sims Causality. *Econometrica*, **50**, 569–582.

Cox, D. R. (1958) *Planning of Experiments.* New York: Wiley.

Dawid, A. P. (1995) Discussion of 'Causal diagrams for empirical research' by J. Pearl. *Biometrika*, **82**, 689–690.

Eerola, M. (1994) *Probabilistic causality in longitudinal studies*, vol. 92. Lecture Notes in Statistics. Berlin: Springer.

Engle, R. F., Hendry, D. F., and Richard, J. F. (1983) Exogeneity. *Econometrica*, **51**, 277–304.

Fisher, R. A. (1925) *Statistical Methods for Research Workers*, 1st edn. Edinburgh: Oliver and Boyd.

Fisher, R. A. (1935) *The Design of Experiments.* New York: Hafner.

Florens, J. P., and Fougere, D. (1996) Noncausality in Continuous Time. *Econometrica*, **64**, 1195–1212.

Florens, J. P., and Mouchart, M. (1982) A Note on Noncausality. *Econometrica*, **50**, 583–592.

Florens, J. P., and Mouchart, M. (1985) A Linear Theory for Noncausality. *Econometrica*, **53**, 157–176.

Gill, R. D., and Johansen, S. (1990) A survey of product-integration with a view towards application in survival analysis. *Ann. Statist.*, **18**, 1501–1555.

Granger, C. W. J. (1969) Investigating Causal Relations by Econometric Models and Cross-spectral Methods. *Econometrica*, **37**, 424–438.

Holland, P. W. (1986) Statistics and Causal Inference (with discussions). *J. Am. Statist. Ass.*, **81**, 945–970.

Hume, D. (1739) *A treatise of human nature.* London: John Noon.

Hume, D. (1748) *An Enquiry Concerning Human Understanding.* London: A. Millar.

Jacod, J. (1975) Multivariate Point Processes: Predictable Projection, Radon-Nikodym Derivatives, Representation of Martingales. *Z. Wahrsch. Ver. Geb.*, **31**, 235–253.

Karr, A. (1991) *Point processes and their statistical inference*, 2nd edn. Marcel Dekker.

Kuhn, T. (1962) *The structure of scientific revolutions.* Chicago: University of Chicago Press.

Mykland, P. (1986). Statistical Causality. *Technical Report 14.* Department of Mathematics, University of Bergen, Norway.

Neyman, J. (1923) On the Application of Probability Theory to Agricultural Experiments. Essay on Principles, Section 9. Translated in *Statist. Sci.*, **5**, 465–480, 1990.

Pearl, J. (1995) Causal diagrams for empirical research (with discussion). *Biometrika*, **82**, 669–710.

Popper, K. (1959) *The Logic of Scientific Discovery. (translation of Logik der Forschung).* London: Hutchinson.

Robins, J. (1986) A new approach to causal inference in mortality studies with a sustained exposure period – application to control of the healthy worker survivor effect. *Math. Model.*, **7**, 1393–1512.

Robins, J. (1989) The control of confounding by intermediate variables. *Statist. in Med.*, **8**, 679–701.

Robins, J., Blevins, D., Ritter, G., and Wulfsohn, M. (1992) G-estimation of the effect of prophylaxis therapy for pneumocystis carinii pneumonia on the survival of AIDS patients. *Epidemiology*, **3**, 319–336.

Robins, J. M. (1997) Marginal structural models. *Proc. Am. Statist. Ass.*, 1–10.

Robins, J. M. (1998) *Structural nested failure time models.* In *Encyclopedia of Biostatistics* (eds Armitage, P. and Colton, T.) Chichester: Wiley & Sons, 4372-4389.

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

Rubin, D. B. (1974) Estimating causal effects of treatments in randomized and nonrandomized studies. *J. Edu. Psych.*, **66**, 688–701.

Rubin, D. B. (1978) Baysian inference for causal effects: the role of randomization. *Ann. Statist.*, **6**, 34–58.

Rubin, D. B. (1991) Practical Implications of Modes of Statistical Inference for Causal Effects and the Critical Role of the Assignment Mechanism. *Biometrics*, **47**, 1213–1234.

Schweder, T. (1970) Composable Markov processes. *J. Appl. Probab.*, **7**, 400–410.

Sims, C. A. (1972) Money, Income and Causality. *Am. Economic Rev.*, **62**, 540–552.

Sosa, E., and Tooley, M. (1993) *Causation.* In Oxford Readings in Philosophy. Oxford: Oxford University Press.

Suppes, P. (1970) *A Probabilistic Theory of Causality.* Amsterdam: North-Holland.