7

Causal inference from observational data: a Bayesian predictive approach

E. Arjas
Department of Mathematics and Statistics, University of Helsinki, Helsinki, Finland

7.1 Background

Causality is a challenging topic for anyone to consider in a formal way. Already the concept itself is problematic, and often people have sharply different opinions of its foundations. However, causality is perhaps a particularly challenging topic for a statistician. Rather than just trying to formulate views on some underlying philosophical issues, a statistician is often faced with the concrete problem of how to find empirical support in favour or against, or even trying to prove or disprove, a causal claim made in some substantive scientific or nonscientific context.

We shall here concentrate on three important aspects of causality. The first aspect is temporality, which manifests itself already in the fundamental requirement that a cause must precede the effect in time. Somewhat surprisingly, the temporal aspect of causality is only rarely explicitly accounted for in the major part of the literature on causal modeling. The second aspect to be emphasized here is confounding. When analyzing observational data with the aim of finding empirical support to a causal claim, there is always a possibility that the differences that are found may in fact be due to spurious associations. While unconfounded inference is ultimately always based on hypotheses that cannot be verified from data, it is important that these hypotheses are formulated in an intuitively understandable and, in the considered context, meaningful way. Third, our approach to causality is completely probabilistic, including aspects

Causality: Statistical Perspectives and Applications, First Edition. Edited by Carlo Berzuini, Philip Dawid and Luisa Bernardinelli.

of statistical inference. This leads us to using predictive distributions as the main summary measures for causal claims.

7.2 A model prototype

To start from a simple setting, suppose we are interested in studying the effect which a contemplated causal variable A might have on some measured response Y. Suppose further that the considered unit, individual or object being studied is described by an observed covariate X. Finally, let U be a generic notation for corresponding unobserved background characteristics and parameters that we think could be relevant to the causal question. When setting up a probability model p for these variables, we consider them in the natural time order $U \rightarrow X \rightarrow A \rightarrow Y$ (meaning U temporally precedes X, which temporally precedes A, and so on), in which case the chain multiplication rule then leads to the joint distribution $p(U, X, A, Y) = p(U)p(X \mid U)p(A \mid U, X)p(Y \mid U, X, A)$. For a graphical representation of this, in the form of a directed acyclic graph (DAG), see Figure 7.1(a).

As the variable U is not observed, predictions concerning Y would be based on only knowing the values of X and A, that is on $p(Y \mid X, A)$. This conditional distribution can be obtained from the joint distribution of all four variables in the obvious way by first forming the marginals $p(X, A, Y)$ and $p(X, A)$ by integration, and then computing $p(X, A, Y)/p(X, A)$.

Here we have particularly in mind the situation arising in purely observational studies, where the investigator has had no real physical control of the value of A even after the value of X was observed. Following a notation introduced by Lindley (2002), this aspect can be emphasized by writing $p(Y \mid X, \text{see}(A))$ for such a prediction. However, for a causal interpretation, where A would be viewed as a cause of Y, it seems essential that one can think of there being a parallel, perhaps only hypothetical, situation in which the value of A would be determined by some external mechanism. This would be the case, for example, in a randomized clinical trial, where A could be the indicator of assigning a patient either to receiving a treatment or to placebo and the external mechanism would be the device used for the randomization. Such a situation is similarly emphasized in the notation, following Pearl (2009), by writing $p(Y \mid X, do(A))$. In what follows, the basis for presenting probabilistic evidence

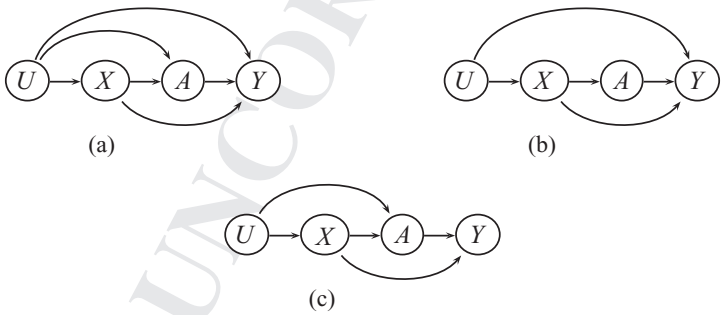

Figure 7.1 A graphical representation of the dependencies considered in the model prototype. In (a) all dependencies may be present and therefore U *is a potential confounder. (b) represents the situation of Definition 1, where* A *is unconfounded relative to* U *and in (c) the model is unconfounded since* U *is not a potential confounder.*

for the claim 'A causes Y' is by considering contrasts between predictive probabilities of the form $p(Y \mid X, do(A = a))$ and $p(Y \mid X, do(A = a'))$, where a and a' are two different values of A that are being compared.

From the perspective of statistical inference, the key question is now the following: 'How can such *do*-probabilities be evaluated empirically when the supporting data come from an observational *see*-study?' The *see*- and *do*-probabilities must obviously be related to each other in some way. However, they are not the same! In particular, the *do*-experiment does not in itself give rise to a specification of a conditional distribution, which would correspond to $p(A \mid U, X)$ in a *see*-study. Indeed, in a *do*-experiment it would be irrelevant for a causal analysis of what random or other mechanism was used to deliver the value of A as long as it was somehow exogenous from the perspective of the observational *see*-study. In this sense one could say that in the *do*-situation the joint distribution of (U, X, A, Y) is only partly specified. In an observational *see*-study, on the other hand, one generally views the variable A as being endogenous, with values being determined in the same fashion as those of the covariate X.

In contrast to this, in a causal context it seems both natural and necessary to assume that otherwise the probabilistic structure and description of the *see*-study is lifted to the, perhaps only hypothetical, *do*-experiment without change (cf. Lindley, 2002). In particular, one should then assume that:

(i) the joint (prior) distribution of the variables U and X is the same in the hypothetical *do*-experiment as in the real *see*-study, and also that

(ii) the response Y, given X, A and U, behaves in the same way regardless of whether the event $\{A = a\}$ was *done* or merely *seen*. Stated more explicitly, we assume that $p(Y \mid U, X, \text{see}(A)) = p(Y \mid U, X, \text{do}(A))$.

The problems therefore seem to centre around the conditional distribution $p(A \mid U, X)$, the potential fallacy for a causal analysis being that of *confounding*: in an observational study it could happen that differences between the predictions $p(Y \mid X, see(A = a))$ and $p(Y \mid X, see(A = a'))$ given to two items or individuals, or to the same item or individual in two different circumstances, with different values a and a' of the contemplated causal variable A, would actually be due, not to these differences in A but to differences in the unobserved variable U. In an extreme situation, A could be only a marker of some important background variable U such that its value alone would completely determine the value of Y. A manipulation that would change only the value of the marker A, and not that of U, would then leave the value of Y unchanged.

This idea can be formalized as follows. Consider the causal problem where X is an observed covariate, A is a contemplated cause and Y the response of interest. Then we call the variable U a *potential confounder* in this causal problem if the prediction of Y based on knowing the values of X, A and U, as expressed by the distribution $p(Y \mid X, A, U)$, actually depends on U. If the considered problem formulation does not contain such potential confounders, then we can say that this causal model is *unconfounded*.

Consider then the more interesting situation where the model contains an unobserved potential confounder U. (If there are more than one, we include them all in U and consider U to be a vector.) Then, as we cannot ignore this variable in our causal analysis, we have to find some condition under which the predictions based on a *see*-study could be utilized also in a situation in which A was thought to arise from a manipulation, that is from a *do*-study.

This becomes possible under the following conditional independence postulate (cf. Arjas and Parner (2004), Definition 1):

Definition 1 *We say that A is unconfounded relative to a potential confounder U if A and U are conditionally independent given X, that is if*

$$p(A \mid U, X) = p(A \mid X)$$

This postulate, which is expressed in the form of a DAG in Figure 7.1(b), has an important consequence. The (posterior) distribution of U based on the observations X and A does not actually depend on A, that is we have:

Lemma 1 *If A is unconfounded in the sense of Definition 1, then the two posterior distributions* $p(U \mid X, see(A))$ *and* $p(U \mid X)$ *are the same.*

This result justifies why it is natural to call A satisfying the condition of Definition 1 unconfounded. The result itself is an obvious consequence of Bayes' formula. For deriving the posterior $p(U \mid X, see(A))$ we would start from the prior $p(U)$ and consider the expression $p(X, A \mid U) = p(X \mid U)p(A \mid U, X) = p(X \mid U)p(A \mid X)$ as the likelihood. Similarly, for deriving $p(U \mid X)$, we would start from the prior $p(U)$ and use $p(X \mid U)$ as the likelihood. However, as the priors are the same and the likelihood expressions are proportional (in U), the posteriors will also be the same. This is because the proportionality constant $p(A \mid X)$ appears in Bayes' formula both in the numerator and the denominator, and therefore can be cancelled.

Now, if A in a *see*-study is unconfounded in the sense of Definition 1, we can replace the rule $p(A \mid X)$ by which the selection of the value of A was modeled, by an arbitrary distribution that does not depend on U. As an extreme case, we could imagine that the values of A observed in the data had in fact been chosen by us in advance, and then fixed. Alternatively, we could think that they had been determined by some rule that had been chosen in advance, as functions of the corresponding covariate values X. The rule could even involve some external information such as independent randomization, as long as it does not depend on the unobserved background variable U. In all such situations, the posterior inference concerning U, when based on X, will remain the same as in the original *see*-study because it does not depend on A.

These considerations become perhaps more clear if we introduce separate notations for the probability distributions describing these two situations, using subscript 'obs' for the observational *see*-study that gave rise to the data and subscript 'ex' for the hypothetical *do*-experiment, this notation referrring to the idea that the selection of the value of the contemplated causal variable A is *exogenous* when viewed from the perspective of a causal model used for describing the observational data. The subscript 'ex' could also be thought of as referring to 'external' or to 'experimental'. (Note that in Arjas and Parner, 2004 the subscript 'opt' has been used in place of 'ex', as an abbreviation of 'optional'. However, such a notation can lead to a confusion between 'optional' and 'optimal'. In some instances we are actually interested in finding an optimal regime for assigning values to A, so we will reserve the subscript 'opt' to denote such a rule; see, for example, Arjas and Saarela, 2010.)

Thus we write, by again applying the chain rule,

$$p_{\text{obs}}(U, X, A, Y) = p_{\text{obs}}(U)p_{\text{obs}}(X \mid U)p_{\text{obs}}(A \mid U, X)p_{\text{obs}}(Y \mid U, X, A) \qquad (7.1)$$

for the joint distribution of the variables (U, X, A, Y) in the data, and then assuming that $p_{\text{obs}}(A \mid U, X) = p_{\text{obs}}(A \mid X)$ if A is unconfounded. In the hypothetical case where the value of A is assigned by some external mechanism, we write similarly

$$p_{\text{ex}}(U, X, A, Y) = p_{\text{obs}}(U)p_{\text{obs}}(X \mid U)p_{\text{ex}}(A \mid U, X)p_{\text{obs}}(Y \mid U, X, A) \tag{7.2}$$

Thus, corresponding to the assumed links (i) and (ii) above, the other parts of the model p_{obs}, apart from the one describing assignment of A, have been retained when switching from 'obs' to 'ex'. Using this notation we can then conclude the following.

Lemma 2 *If* A *is unconfounded in the sense of Definition 1, then:*

(i) The posterior distributions of U based on the observed data (X, A) *are the same in both schemes, that is*

$$p_{\text{obs}}(U \mid X, A) = p_{\text{ex}}(U \mid X, A) \tag{7.3}$$

Here neither of these posterior distributions depends on A.

(ii) The posterior distributions of U based on the observed data (X, A, Y) *are the same in both schemes, that is*

$$p_{\text{obs}}(U \mid X, A, Y) = p_{\text{ex}}(U \mid X, A, Y) \tag{7.4}$$

(iii) The predictive distributions of Y based on observed data (X, A) *are the same in both schemes, that is*

$$p_{\text{obs}}(Y \mid X, A) = p_{\text{ex}}(Y \mid X, A) \tag{7.5}$$

This first claim is a restatement of Lemma 1. The second claim follows directly from the first, by an application of Bayes' formula, when interpreting $p_{\text{obs}}(U \mid X, A)$ and $p_{\text{ex}}(U \mid X, A)$ in (7.3) as prior distributions and then using the assumption that the respective conditional distributions (or likelihood expressions) $p_{\text{obs}}(Y \mid U, X, A)$ and $p_{\text{ex}}(Y \mid U, X, A)$ of Y are the same in both schemes. Proving the third conclusion is similar; it follows at once from (7.3) when multiplying the left-hand side by $p_{\text{obs}}(Y \mid U, X, A)$ and the right-hand side by $p_{\text{ex}}(Y \mid U, X, A)$, which factors were assumed to be identical, and then integrating U out on both sides. Note that the same conclusion (7.5) holds trivially if U is not a potential confounder, that is, $p_{\text{obs}}(Y \mid U, X, A) = p_{\text{obs}}(Y \mid X, A)$, a situation displayed graphically in Figure 7.1(c).

Stated briefly, the first two results say that under the unconfoundedness condition the posterior inferences concerning U when based on X and A, and similarly when based on X, A and Y, are the same regardless of whether the value of A has been *done* or merely *seen*. As a consequence, under that condition, we can perform Bayesian estimation of the unknown variable U in the context of an observational study as if treatment A had been assigned, in a designed experiment, some pre-specified fixed value.

Essentially the same conclusion as in (7.4) holds if, instead of applying Bayesian methods and terminology, one prefers using direct likelihood inference and a corresponding

formulation of the result. Then, fixing X and considering $p_{\text{obs}}(A, Y \mid U, X) = p_{\text{obs}}(A \mid U, X)p_{\text{obs}}(Y \mid U, X, A)$, as a function of U, to be the likelihood arising from observing A and Y, we see readily from the unconfoundedness postulate that its first factor does not depend on U and that it is therefore proportional (in U) to its second factor $p_{\text{obs}}(Y \mid U, X, A)$. Thus only this second factor, which is common to both schemes, needs to be considered when inferring on U.

Remark 1 The above simple model is in many ways similar to the well-known Rubin causal model (Rubin, 1974; Holland, 1986). Both models can be viewed as providing descriptions of a situation in which, on a given individual i, covariates X_i are measured at time 0 and this is immediately followed by an assignment A_i of i to either treatment or placebo. At a later point in time τ, a response Y_i is measured. In the Rubin causal model, however, it is assumed that for each individual i there exist, already before the experiment is performed, two fixed *potential outcomes*, $Y_0(i)$ under placebo and $Y_1(i)$ under treatment (for a detailed discussion of the concept of potential outcomes, see Chapter 2 in this volume by Sjölander). The treatment assignment A_i then determines which potential outcome will actually be realized and observed (and only one will). The randomness in the Rubin model is thus contained in the assignment variables A_i for different individuals i. In the terminology of Rubin, and now suppressing the index i in the notation, A is called *ignorable* if it is conditionally independent of (Y_0, Y_1) given X, that is $(Y_0, Y_1) \perp A \mid X$. This corresponds closely to our Definition 1 of unconfounded inference, and actually formally coincides with it if we set $U = (Y_0, Y_1)$. Note also that if this connection between the two models is made, U becomes a trivial potential confounder in our setting because the value of Y is completely determined by U and A. The main difference between the Rubin model and ours is in the interpretation of U and, more generally, in how randomness is understood. A technical difference is that in the Rubin model separate variables and notations are introduced for different potential outcomes, whereas in our approach there is a single response variable Y and we consider its conditional distributions given A.

7.3 Extension to sequential regimes

These simple considerations extend readily to a longitudinal setting in which we have in the data a sequence of event times and corresponding descriptions of the events that have taken place. Such a setting is of particular interest in clinical studies, where one attempts to establish an optimal regime for the treatments and where the individual treatment decisions are allowed to depend on how a patient has responded to the treatments that were received earlier. Here is a general formulation of such a situation.

Suppose that a random number N_τ events occur over the considered time interval $(0, \tau]$. At each event time T_k, covariates X_k are measured and an action, or treatment, A_k follows immediately upon this. Hence, the recorded data consist of $(T_k, (X_k, A_k))$, $k = 0, 1, \ldots, N_\tau$, with $0 \equiv T_0 < T_1 < T_2 < \cdots < T_{N_\tau}$, and finally, of a measured response Y. Such a sampling scheme can be formulated naturally by using the general framework of marked point processes (see, for example, Brémaud, 1981; Arjas, 1989; Karr, 1991; Andersen *et al*., 1993), with T_k being the event times and $Z_k = (X_k, A_k)$ the corresponding marks. As a convention, we can also treat the considered response Y as a marked point, identifying it with 'the last observed covariate value' X_{N_τ}. This can be done without any restriction to the generality and allows, for example, situations in which the final response is some summary measure determined on the basis of the entire observed sample path of the process.

The marked point process formalism is also able to accommodate latent variables and developments, which are potentially relevant for describing the considered causal problem, but which were not observed. As before, we use the generic notation U for such variables and then make the convention that they can be imbedded, as a sequence of additional random marks, into the marked point process. Having denoted by T_k the time of the kth event, we denote by $\tilde{Z}_k = (U_k, X_k, A_k)$ the corresponding event, or mark, where (X_k, A_k) is observed in the data and U_k is unobserved.

Before moving on, we make some further conventions on how the marks can be interpreted. In a situation in which we have data on several individuals, say n individuals indexed by $i = 1, 2, \ldots, n$, we can form the natural superposition of the marked point processes describing the considered individuals, then arriving at a formulation in which the components of $\tilde{Z}_k = (U_k, X_k, A_k)$ are vectors with coordinates indexed according to the individuals. In some designs, e.g. randomized clinical trials, there may not be a covariate measurement X_k preceding a corresponding assignment A_k to a treatment, say a, in which case we could write $(\varnothing, a)$ as the value of (X_k, A_k). On the other hand, in some sampling schemes a number of repeated covariate measurements are made before there is an actual assignment A_k to a treatment, and then we can similarly write $(x, \varnothing)$. A more general but also notationally more elaborate description and analysis of such situations is considered in Parner and Arjas (1999).

In problems of this type it seems natural to think that there is some structural interest parameter θ, which is common to all these individuals and with which the inferential problem is primarily concerned. On a more abstract level, the existence of such a parameter would be justified by an exchangeability assumption and the de Finetti (1974) representation theorem. We use θ as a generic notation for model parameters; its exact meaning will then depend on the particular application and model that are considered. We now make the convention that θ is imbedded into the latent mark U_0 as a coordinate. Thus our inferences concerning the latent marks $U_k, k = 0, 1, \ldots, N_\tau$, will also cover inferences on θ.

Setting up a probability for the canonical sample paths of such a marked point process turns out to be very simple because the marked point process framework allows us to proceed by induction, always moving from a time point T_k to the next point at T_{k+1} and then considering it jointly with the corresponding mark $(U_{k+1}, X_{k+1}, A_{k+1})$. Setting up a probability model for the sample paths of the marked point process can then be done by a sequential application of the chain rule of multiplication. In its kth step we consider conditional probabilities of the form $p_{\text{obs}}(T_{k+1}, U_{k+1}, X_{k+1}, A_{k+1} \mid \mathcal{F}_k)$, where $\mathcal{F}_k = \{(T_i, U_i, X_i, A_i); i = 0, 1, \ldots, k\}$ is the 'full' history of the marked point process up to time T_k. In this way it will be sufficient to consider a single generic step in such an induction. Denote similarly by $\mathcal{H}_k = \{(T_i, X_i, A_i), i = 0, 1, \ldots, k\}$ the observed history up to time T_k.

Consider then the issue of potential confounding in this framework. We start with the following definition, which extends our earlier Definition 1 to the present sequential setting and corresponds to Definition 2 in Arjas and Parner (2004) (see also Section 8.8 in Chapter 8 in this volume by Berzuini, Dawid and Didelez and Definition 2 in Dawid and Didelez 2010).

Definition 2 Unconfounded inference *We say that a sequence of contemplated causal variables (A_k) in an observational study described by sample path $\mathcal{F}_{N_\tau} = \{(T_i, U_i, X_i, A_i), i = 0, 1, \ldots, N_\tau\}$ and probability p_{obs} is unconfounded relative to latent variables* (U$_{\text{k}}$) *if, for each* k, *A_k and $\{U_i, i = 0, 1, \ldots, k\}$ are conditionally independent given $(\mathcal{H}_{k-1}, T_k, X_k)$,*

that is

$$p_{\text{obs}}(A_k|\mathcal{F}_{k-1}, T_k, U_k, X_k) = p_{\text{obs}}(A_k|\mathcal{H}_{k-1}, T_k, X_k), k = 1, 2, \ldots, N_\tau \tag{7.6}$$

Remark 2 The above setting bears a close resemblance to the sequential trial design studied by Robins (1986). In particular, our condition of unconfounded inference corresponds to his 'no unmeasured confounders' condition. Robins' model can be viewed as a extension of the Rubin causal model to sequential designs, and its main technical – and perhaps also conceptual – difference to ours is its reliance on the potential outcomes framework. More specifically, Robins considers potential outcomes of the form $Y_{\mathbf{a}}$, where $\mathbf{a} = (a_0, \ldots, a_K)$ is a regime of a fixed sequence of K treatments at predetermined time points. Its interpretation is similar to the Rubin causal model, in the sense that the potential outcomes $Y_{\mathbf{a}}$ are then assumed to exist, in some sense, already before the experiment has been performed, for all admissible regimes $\mathbf{a}$.

Let us now see how this postulate can be used in a context where we would like to relate an observed sequence of events to a causal problem, viewing the observed variables (A_k) as 'causes'. Following the same idea as in Section 7.2, we connect the inferences, which can be drawn from the observational data and which are described in terms of a probability denoted by p_{obs}, to corresponding statements relative to another probability denoted by p_{ex}. To do so, we link these two probabilities to each other by the following requirements:

$$\begin{gathered} p_{\text{ex}}(U_0, X_0) = p_{\text{obs}}(U_0, X_0) \\ p_{\text{ex}}(T_{k+1}, U_{k+1}, X_{k+1} \mid \mathcal{F}_k) = p_{\text{obs}}(T_{k+1}, U_{k+1}, X_{k+1} \mid \mathcal{F}_k),\ k = 0, 1, \ldots, N_\tau \end{gathered} \tag{7.7}$$

Note that again in here, we can view the probability p_{ex} as being only partially defined in the sense that the conclusions of Theorem 1 below remain valid regardless of the way in which the treatment assignment probabilities for (A_k) are specified, as long as they are exogenous in the sense that they do not depend on the potential confounder variables (U_k).

Our main conclusion can now be stated as follows.

Theorem 1 *Suppose that contemplated causal variables (A_k) are unconfounded in the sense of Definition 2. Then, for each $k = 0, 1, \ldots, N_\tau$:*

(i) The posterior distributions of the complete history $\mathcal{F}_k = \{T_i, U_i, X_i, A_i; i = 0, 1, \ldots, k\}$, given the corresponding observed history $\mathcal{H}_k = \{T_i, X_i, A_i; i = 0, 1, \ldots, k\}$, are the same in both schemes, that is

$$p_{\text{obs}}(\mathcal{F}_k|\mathcal{H}_k) = p_{\text{ex}}(\mathcal{F}_k|\mathcal{H}_k) \tag{7.8}$$

Here neither of these posterior distributions depends on the latest assignment A_k.

(ii) The predictive distributions of the next marked point $(T_{k+1}, U_{k+1}, X_{k+1})$ given the corresponding observed history $\mathcal{H}_k = \{T_i, X_i, A_i, i = 0, 1, \ldots, k\}$ are the same in both schemes, that is

$$p_{\text{ex}}(T_{k+1}, U_{k+1}, X_{k+1} \mid \mathcal{H}_k) = p_{\text{obs}}(T_{k+1}, U_{k+1}, X_{k+1} \mid \mathcal{H}_k) \tag{7.9}$$

For a proof of (7.8), we only need to go through the steps of Lemma 2, replacing (U, X, A) with $(U_i, (T_i, X_i), A_i)_{0 \leqslant i \leqslant k}$, and use induction in k. The crucial argument in such an induction is that, when moving from T_k to T_{k+1}, the likelihood contribution corresponding to the 'new data' $(T_{k+1}, X_{k+1}, A_{k+1})$ is, possibly up to a proportionality factor not depending on $(U_i)_{0 \leqslant i \leqslant k+1}$, the same as what would correspond to only (T_{k+1}, X_{k+1}). To prove (7.9), note first that it states the same relationship between the conditional distributions of $(T_{k+1}, U_{k+1}, X_{k+1})$ as (7.7), except that now the conditioning is with respect to the observed history $\mathcal{H}_k$. However, this follows from (7.8) when the transition probabilities (7.7), considered as functions of U_k, are integrated out with respect to the corresponding (identical) posterior distributions p_{ex} and p_{obs}.

Remark 3 This same result could have been derived and presented by using the framework of counting processes and the associated stochastic intensities (more generally, compensators) based on a continuous time parameter $t \geqslant 0$. The postulate of unconfounded inference, corresponding to our Definition 2, would then be phrased in a natural way as a *local independence* condition (Schweder, 1970; Aalen, 1987; Didelez, 2008), technically stating that the local characteristics corresponding to treatment assignments in the compensators would be the same, regardless of whether the compensators were considered to be relative to the observed histories $(\mathcal{H}_t)_{t \geqslant 0}$ or relative to the larger histories $(\mathcal{F}_t)_{t \geqslant 0}$ generated, in addition, by the unobserved process $(U_t)_{t \geqslant 0}$. (Note that, in the somewhat different context of time series, the concept of *noncausality* of Granger (1969) expresses a very similar idea as local independence.)

These two ways of formulating the key condition, one based on considering a sequence of marked points and the associated conditional distributions as above, and the other on the corresponding counting processes in continuous time and their compensators with respect to alternative histories, are easily seen to be equivalent. This is so because, in a marked point process, the compensators between any two consecutive event time points T_k and T_{k+1} are actually completely determined by the conditional distributions of $(T_{k+1}, U_{k+1}, X_{k+1}, A_{k+1})$ given the respective histories $\mathcal{H}_k$ and $\mathcal{F}_k$ at T_k, and conversely. More generally, both these approaches can be viewed to be special cases of *filtering* of a marked point process (see, for example, Arjas *et al*., 1992; Arjas and Haara, 1992).

The approach based on continuous time and a local independence condition has been considered in complete detail in Parner and Arjas (1999). However, instead of considering posterior and predictive distributions as in Theorem 1, the main result there is formulated in terms of likelihood processes. It says that, under the local independence condition, the contributions from the treatment assignments A_k to the likelihood expression do not depend on the parameter of interest. These contributions can therefore be ignored in direct likelihood inference concerning that parameter. Thus the comments on likelihood inference that were made earlier, after Lemma 2, remain valid also in this more general context of sequential assignments.

Remark 4 There is an interesting difference in how knowledge of A_k influences conditional probability distributions describing the *past* and the *future*. While, as stated in part (i) of Theorem 1, the posterior distribution of the latent history $\{U_i; i = 0, 1, \ldots, k\}$, based on $\mathcal{H}_k = \{T_i, X_i, A_i; i = 0, 1, \ldots, k\}$, does not depend on A_k, all future variables in the marked point process after time T_k, including the final response, may well depend on it. Indeed, such dependence is precisely what a statistician anticipating a causal effect would expect to see.

Note also that a similar statement does not hold for covariates X_k; a new covariate reading will generally have both inferential value backwards in time and predictive value forwards in time.

7.4 Providing a causal interpretation: predictive inference from data

Looking now at the contents of Theorem 1 one may wonder what bearing it has for considering concrete problems in causality. While it provides a condition under which unconfounded statistical inferences can be drawn from observational data, it does not at first sight appear to say how such inferences could be used to arrive at useful conclusions relating to causality. In the following we try to straighten this, before following the same basic ideas that were, in a simple case, presented in Section 7.2.

Suppose that we, after having carried out data analysis of the above kind, would like to express our conclusions by using 'causal language'. For such a purpose it would be natural to try to predict how a generic (real or hypothetical) item or individual would respond to some specific treatment, or a sequence of such treatments. Using from now on the word 'individual' in this case, we then assume that (i) the contemplated causal variables (A_k) considered in the data are unconfounded in the sense of Definition 2 and (ii) this generic individual is exchangeable with those considered in the data in the sense that their behavior can be modeled by the same probability model, parametrized with a common parameter θ. In other words, we assume that the inferences drawn from the observational data are valid information about how a similar individual would respond to treatments assigned by some exogenous rule or mechanism.

The inferences that were drawn from the data $\mathcal{H}_{N_\tau} = \{T_k, (X_k, A_k), k = 0, 1, \ldots, N_\tau\}$ (henceforth denoted simply by 'data'), on the common model parameter θ can then be utilized also for predicting what will happen to this generic individual if he/she is to be given some specific sequence of treatments. Adding a star (*) to the notation to signify the considered generic individual, possibly described by some covariate values $X_0^* = x_0^*$ at the baseline, we would, in a simple case, be interested in predicting the response Y^* under a given fixed sequence of 'forced' treatment assignments, say $A_i^* = a_i^*, i = 0, 1, \ldots, k$ (cf. Dawid, 2002). Such a prediction can then be expressed in terms of the predictive distribution $p_{\text{ex}}(Y^* \mid x_0^*, \mathbf{a}^*, \text{data})$, where we have denoted $\mathbf{a}^* = (a_i{}^*)_{0 \leqslant i \leqslant k}$.

More generally, there could be a *dynamic treatment regime*, say $\mathcal{A}$, such that each A_k^* could be allowed to be a function of the past observed history of that individual, consisting of past event times, covariate readings and possible earlier treatment assignments. In fact, the treatment assignments of an individual could in principle depend also on the past histories of the other individuals in the sample. Thus we are not, generally speaking, postulating here a 'no interference between units' property. More generally still, such a regime could be randomized, as long as the randomization mechanism does not depend on the past potential confounder variables U_k^*. In order to make the role of the regime $\mathcal{A}$ explicit in the notation we add it to the subscript of p_{ex} by writing $p_{\text{ex}(\mathcal{A})}$. The considered predictive distribution will then be denoted by $p_{\text{ex}(\mathcal{A})}(Y^* \mid x_0^*, \text{data})$. The exact specification of this probability will then depend on the considered assignment mechanism $\mathcal{A}$.

Note that before finally obtaining a response Y^*, the considered generic individual is thought to experience a realization of the marked point process; that is there may be a

sequence of event times T_k^*, covariate readings X_k^* and assignments A_k^* according to the chosen regime $\mathcal{A}$. Apart from possible fixed attributes (such as gender), which have been initially chosen to characterize the considered individual, the covariates will generally evolve over time in ways that cannot be known precisely at the time at which the prediction is made. As a consequence, the computation of these predictive distributions involves an integration in the assumed probability model with respect to such possibly time-dependent random variables.

It may be useful to distinguish between three different ingredients, or sources of randomness, which need to be accounted for when computing the predictive distributions $p_{\text{ex}(\mathcal{A})}(Y^* \mid x_0^*, \text{data})$. The first is the intrinsic or endogenous randomness in the behavior of the considered generic individual, that is in the event times $T_k{}^*$, covariate readings $X_k{}^*$ and ultimately response Y^*, as specified by our probability model p_{ex} in a situation in which the values of θ, X_0^* and $\mathbf{A}^*$ would be fixed at some known values. The second potential source is the treatment regime $\mathcal{A}$, which may be randomized by some mechanism. Finally, the model parameters θ are unknown and need to be estimated from data. Here, using our earlier convention that the parameters θ are already contained in the common latent variable U_0, which belongs to $\mathcal{F}_0$, the necessary results are summarized by the corresponding marginal of the posterior distribution $p_{\text{obs}}(\mathcal{F}_{N_\tau} \mid \text{data})$, which was provided in Theorem 1 (i). We can then compute, by an integration with respect to this posterior, the desired predictive distribution $p_{\text{ex}(\mathcal{A})}(Y^* \mid x_0^*, \text{data})$.

Having now described how to derive, for a given regime $\mathcal{A}$, the predictive distribution $p_{\text{ex}(\mathcal{A})}(Y^* \mid x_0^*, \text{data})$ of the generic response Y^*, given the values of the baseline covariates, we can consider any two such regimes of interest, say $\mathcal{A}_1$ and $\mathcal{A}_2$, and compare the corresponding predictive distributions $p_{\text{ex}(\mathcal{A}_1)}(Y^* \mid x_0^*, \text{data})$ and $p_{\text{ex}(\mathcal{A}_2)}(Y^* \mid x_0^*, \text{data})$, or the corresponding expectations of some suitable test functions, to each other. In practice, the necessary numerical integration can be carried out efficiently by Monte Carlo simulation, by applying data augmentation alongside the computations that are needed for statistical inference (see also Section 8.10 in Chapter 8 in this volume by Berzuini, Dawid and Didelez).

Practical illustrations of this general method can be found, for example, in Arjas and Liu (1995, 1996), Arjas and Haastrup (1996), Arjas and Andreev (2000), Arjas and Saarela (2010) and Härkänen *et al*. (2000).

Remark 5 The above predictive distributions bear a close resemblance to the G-computation algorithm of Robins (1986) (see also Dawid and Didelez, 2010 and Parner and Arjas, 1999, for some comments, the latter being available on the web page http://wiki.helsinki.fi/display/biometry/Arjas).

Technically, there seems to be the difference that our predictive distributions involve additional integrations, not only of time points (T_k^*) and (potentially time dependent) future covariates (X_k^*) but also of the model parameter θ (with respect to the posterior) and of future treatment assignments (with respect to the considered regime $\mathcal{A}$). The expression obtained by the G-computation algorithm thus, generally speaking, depends on the model parameter and a chosen sequence of treatment assignments, where the latter could be said to represent a sequence of 'forced conditionings' (instead of standard conditioning of probabilities). Our use of predictive distributions, on the other hand, integrates both the results from statistical inference (represented by the posterior) and adherence to a considered dynamic regime into the same computation. Another difference is more conceptual, stemming from the fact that

we are here not making use of formulations based on the idea of potential or counterfactual outcomes. This aspect seems to separate us from a large body of the causality literature.

7.5 Discussion

The main issue in this paper has been our attempt to clarify how, and under what circumstances, unconfounded statistical inferences can be drawn from observational data by applying Bayesian methods and modeling based on marked point processes. It was argued that the causal conclusions from such study, in the sense of effects of causes (Dawid, 1995), can be formulated in a natural way in terms of posterior predictive distributions or the corresponding expectations. Such results can then be seen as ingredients in the lengthy process of arriving at a causal understanding and interpretation of the empirical results obtained in some substantive context of interest (Cox, Chapter 1 in this volume). However, since our approach is entirely probabilistic, it becomes important how the concept of probability is understood in such a context. Our point of view has been that probabilities should be viewed primarily, not as objects or characteristics of the physical world but as expressions of what is known about it on the basis of the information that is available. Consistent with this, we are here not claiming to be providing tools for proving the existence, in a literal sense, of some causal relationships in Nature, nor for identifying the magnitudes of corresponding effects.

Considering briefly two well-known popular alternatives to the approach presented here, graphical models provide a convenient tool for exploring and visualizing the dependence relations between random variables that are involved (e.g. Pearl, 2009; Didelez, 2008). This is especially true for conditional independencies since they can be read off directly from the graph. Although directed acyclic graphs (DAGs) often reflect the time order in which sampling on a single individual was performed, they do not fully incorporate the time aspect. More specifically, they do not specify *when* in time measurements and actions were taken. This can be crucial if inference is based on a sample of individuals and sampling does not follow a sequential trial design with predetermined time points. Moreover, graphical representations of such situations tend to become impracticable enough to lose their intuitive appeal and usefulness for visualization. (Some additional aspects relating to the role of the time in causal considerations are provided in Arjas and Eerola, 1993.) Finally, one sees only rarely explicit suggestions on how statistical inferences drawn from existing empirical data should be used for assigning probabilities to a DAG.

The second popular framework for considering causal modeling and inference, based on the concept of potential or counterfactual outcomes, may be useful because of the intuitive interpretations they offer. However, their drawback, particularly in the context of sequential regimes, is the large number of random variables representing such potential outcomes 'that might have occurred' and the consequent tedious notation. One may also wonder how much sense the modeling convention, according to which all these variables are 'fixed', makes as a description of reality. Marked point processes, the approach advocated here, seem to be more flexible in these respects while also allowing for natural extensions.

It is generally accepted that the randomization 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 randomization managed to distribute the unobserved characteristics over treatment groups. This

can be a problem, particularly in small samples. Thus, while the conditions for unconfounded inference, given in Definitions 1 and 2, are guaranteed to hold in designs in which treatment assignment is through a genuine randomization device, this will only protect us against biasing our causal conclusions systematically by unwarranted confounding. In concrete applications, randomization is therefore a valid argument for claiming that the analysis is not confounded, but in no way an absolute safeguard.

In practice, suggesting candidates for the marks (U_k, X_k, A_k) is limited to the *causal field* under consideration, that is to observed and unobserved variables and processes that the investigator is conjecturing, or knows, to be relevant to the considered causal problem. Any statement about causal dependence would then be relative to that setting. 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 'unconfounded inference' assumption of Definition 2.

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 is unconfounded. This may, however, be a dangerous policy because conditioning on additional random variables may destroy an already existing conditional independence property. This is also the reason why, when computing predictive distributions of the form $p_{\mathrm{ex}(\mathcal{A})}(Y^* \mid x_0^*, \text{data})$, we are not controlling the values of the covariates X_k^*, except for $X_0^* = x_0^*$ determined already at the baseline, or those of the later assignments A_k^*, but integrate them out, the latter according to the chosen regime $\mathcal{A}$. This is so for two reasons: first, such variables are often intermediate in the sense of being in time between a contemplated causal variable and the considered response, and can sometimes be viewed as mediators of the causal effect of the former on the latter, and, second, because such conditioning can influence the predicted response also indirectly, by feeding back to the distribution of the unobservables in the past. This is also reflected in our notation of the predictive distributions when we write $p_{\mathrm{ex}(\mathcal{A})}(Y^* \mid x_0^*, \text{data})$, instead of using standard conditioning of the form $p_{\mathrm{ex}}(Y^* \mid x_0^*, \mathbf{A}^*, \text{data})$, where $\mathbf{A}^* = (A_i{}^*)_{0 \leqslant i \leqslant k}$ would be some sequence of treatment assignments.

Acknowledgement

Parts of this article have been taken, after some modification, from Parner and Arjas (1999) and Arjas and Parner (2004).

References

Aalen, O.O. (1987) Dynamic modelling and causality. *Scandinavian Acta Journal*, pp. 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. *Scandinavian Journal of Statistics*, **16**, 177–225.

Arjas, E. and Andreev, A. (2000) Predictive inference, causal reasoning, and model assessment in nonparametric Bayesian analysis: a case study. *Lifetime Data Analysis*, **6**, 187–205.

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

Arjas, E. and Haara, P. (1992) Observation scheme and likelihood. *Scandinavian Journal of Statistics*, **19**, 111–132.

Arjas, E. and Haastrup, S. (1996) Claims reserving in continuous time; a nonparametric bayesian approach. *A.S.T.I.N. Bulletin*, **26**.

Arjas, E. and Liu, L. (1995) Assessing the losses caused by an industrial intervention: a hierarchical bayesian approach. *Applied Statistics*, **44**, 357–368.

Arjas, E. and Liu, L. (1996) Nonparametric Bayesian approach to hazard regression: a case study involving a large number of missing covariate values. *Statistics in Medicine*, **15**, 1757–1770.

Arjas, E. and Parner, J. (2004) Causal reasoning from longitudinal data. *Scandinavian Journal of Statistics*, **31**, 171–187.

Arjas, E. and Saarela, O. (2010) Optimal dynamic regimes: presenting a case for predictive inference. *The International Journal of Biostatistics*, **6**, Article 10 (electronic). DOI: 10.2202/1557-4679.1204.

Arjas, E., Haara, P. and Norros, I. (1992) Filtering the histories of a partially observed marked point process. *Stochastic Processes and their Applications*, **40**, 225–250.

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

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

Dawid, A.P. (2002) Influence diagrams for causal models and inference. *International Statistical Review*, **70**, 161–189.

Dawid, A.P. and Didelez, V. (2010) Identifying the consequences of dynamic treatment regimes: a decision theoretic overview. *Statistics Surveys*, **4**, 184–231.

de Finetti, B. (1974) Bayesianism: its unifying role for both the foundations and applications of statistics. *International Statistical Review*, **42**, 117–130.

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

Granger, C.W.J. (1969) Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, **37**, 424–438.

Härkänen, T., Virtanen, J. and Arjas, E. (2000) Caries on permanent teeth: a nonparametric bayesian analysis. *Scandinavian Journal of Statistics*, **27**, 577–588.

Holland, P.W. (1986) Statistics and causal inference, with discussions. *Journal of the American Statistical Association*, **81**, 945–970.

Karr, A. (1991) *Point Processes and Their Statistical Inference*, 2nd edition. Marcel Dekker.

Lindley, D. (2002) Seeing and doing: the concept of causation. *International Statistical Review*, **70**, 191–214.

Parner, J. and Arjas, E. (1999) Causal reasoning from longitudinal data. Research Report A27, Rolf Nevanlinna Institute, Helsinki.

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

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. *Mathematical Modelling*, **7**, 1393–1512.

Rubin, D.B. (1974) Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, **66**, 688–701.

Schweder, T. (1970) Composable Markov processes. *Journal of Applied Probability*, **7**, 400–410.