 Vol 31: 171–187, 2004

Causal Reasoning from Longitudinal Data*

ELJA ARJAS
University of Helsinki

JAN PARNER
University of Copenhagen

ABSTRACT. This paper reviews some of the key statistical ideas that are encountered when trying to find empirical support to causal interpretations and conclusions, by applying statistical methods on experimental or observational longitudinal data. In such data, typically a collection of individuals are followed over time, then each one has registered a sequence of covariate measurements along with values of control variables that in the analysis are to be interpreted as causes, and finally the individual outcomes or responses are reported. Particular attention is given to the potentially important problem of confounding. We provide conditions under which, at least in principle, unconfounded estimation of the causal effects can be accomplished. Our approach for dealing with causal problems is entirely probabilistic, and we apply Bayesian ideas and techniques to deal with the corresponding statistical inference. In particular, we use the general framework of marked point processes for setting up the probability models, and consider posterior predictive distributions as providing the natural summary measures for assessing the causal effects. We also draw connections to relevant recent work in this area, notably to Judea Pearl's formulations based on graphical models and his calculus of so-called *do*-probabilities. Two examples illustrating different aspects of causal reasoning are discussed in detail.

Key words: Bayesian inference, conditional independence, confounding, marked point processes, predictive distributions

1. Introduction: from association to causation

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. But 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 prove or disprove, a causal claim made in some substantive scientific or non-scientific context.

A common solution is to altogether avoid using the terminology that would refer to causality. Most textbooks on elementary statistics first give the warning that causality and association are not the same, and then go on to discuss only association. For an interesting historical account, see Freedman (1999). Yet one cannot dispute the fact that nearly always, if one tests the hypothesis that a correlation or regression coefficient is zero against the alternative, a small p-value is interpreted as evidence, or even as a proof, that one of the considered variables is in some sense a cause of the other. From a practical point of view, the important question is whether statistics has something better to offer than this.

The philosophical aspects of causality can be traced back to Hume (1739) with his legacy that proof is impossible in empirical science. A very different view has been presented recently in the excellent monograph of Pearl (2000), which, in addition to presenting a new theory of

*This paper was presented by Elja Arjas as a Specially Invited Paper at the 19th Nordic Conference on Mathematical Statistics, Stockholm, June 2002 (NORDSTAT 2002).

causality based on graphical models, also contains a very readable survey of the relevant literature.

Among statisticians, several major contributions have been made over the last couple of decades. In the field of medical statistics a framework called the Rubin causal model (Rubin, 1974, 1978; Holland, 1986), based upon ideas of Neyman (1923), has gained popularity. The key concept here is that of *potential outcome* Y_{ik}, the response of individual i when applying treatment k. The potential outcomes are assumed to be *a priori* fixed for given i and k, and the randomness in the experiment is supposed to arise from how different treatments are assigned to different individuals, resembling the idea behind sample surveys. What Holland (1986) called the Fundamental Problem of Causal Inference is the fact that only one treatment can be applied on a particular individual at a time, thus making causal effects on the unit level unidentifiable from the observed data. Examples of such unit causal effects are measures such as $Y_{i1} - Y_{i0}$ and Y_{i1}/Y_{i0} for the effect of applying on individual i active treatment instead of a placebo. Rubin chose to focus on the average (over individuals) causal effect, which is estimable under the *ignorable treatment assignment* assumption (Rubin, 1978; Rosenbaum & Rubin, 1983). For a two-armed clinical trial this key assumption of valid causal inference is stated as the requirement that $p(A = a \mid X, Y_0, Y_1) = p(A = a \mid X)$ whenever $p(A = a \mid X) > 0$. Here A is an indicator variable telling which of the two possible treatments is chosen, X denotes pre-treatment measured covariates, and Y_0 and Y_1 are the population level potential outcomes under, respectively, placebo and active treatment. In other words, given the covariate X, treatment choice A is assumed to be conditionally independent of Y_0 and Y_1. 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 (and is independent of what such outcomes, if they were known, would be). Robins (1986, 1997) extended this postulate to sequential designs, calling it *no unobserved confounders*.

Although both postulates seem to be relatively straightforward conditional independence statements, it is nevertheless quite hard to understand intuitively what independence from a collection of such potential outcomes would really mean. In particular, this assumption appears to require that the values of the individual potential outcomes would already exist before the action was taken, which does not seem natural in view of the axiomatic time ordering between cause and effect. Considering the above-mentioned Fundamental Problem of Causality, one could even say that their joint existence is only a convention in the modelling that does not correspond to our common perception of the real world. In this sense potential outcomes have some resemblance to the use of a collection of latent lifetimes in competing risks theory, with one latent lifetime for each potential cause of death. There is a single death, however.

Closely related to the idea of the strongly ignorable treatment assignment is the calculus of the *do*-probabilities found in Pearl (1995, 2000). Essentially the idea appeared already somewhat earlier in Spirtes *et al.* (1993). On the subset of structural models that can be represented as directed acyclic graphs (DAGs), Pearl discusses the identification of such *do*-probabilities from the observed data. Like the Rubin causal model, Pearl's framework does not explicitly incorporate time, although the direction of the vertices of the DAGs will generally correspond to the time ordering of variables.

The main points in this paper can be summarized as follows:

- It is important to make the time aspect of causality explicit, not only because 'a cause has to precede the effect' but also because often durations between different events are an integral part of the causal problem, and therefore of the analysis.

- The *ignorable treatment assignment/no unobserved confounders* postulate based on a potential outcomes framework is here replaced by – what we think is more intuitive – consideration of unobserved potential confounder variables/processes, leading to natural conditions under which unconfounded statistical inferences on the causal effects can be drawn.
- While Pearl's *do*-calculus is undoubtedly a useful concept for explaining why interventions should be separated from ordinary conditioning, we nevertheless prefer corresponding statements based on 'ordinary' conditional probabilities (which will be valid under natural conditional independence statements). However, we introduce separate probability measures for the two types of design, making a distinction between designs involving interventions and schemes that are purely observational, but then linking these two probabilities by a set of natural conditions concerning 'the natural evolution' of the processes of interest.
- Bayesian statistical inference based on experimental or observational data can be usefully combined with considering predictive distributions as the main criterion for comparing different treatments in terms of their causal effects. This leads to an 'integrated causal analysis', where a single probabilistic framework is applied both for the statistical inference and for the comparison of treatment effects.

2. Seeing and doing: some preliminary considerations

To start from a simple setting, let X be an observed covariate describing a generic individual, A a contemplated causal variable, Y the observed response, and U some unobserved fixed characteristic that is relevant in the causal analysis. The time ordering in this case is $U \rightarrow X \rightarrow A \rightarrow Y$ so that it is natural to apply the chain multiplication rule for the joint distribution for (U, X, A, Y) in the order $p(u, x, a, y) = p(u)p(x \mid u)p(a \mid u, x)p(y \mid u, x, a)$. As the variable U is not observed, predictions concerning Y would be based on knowing only the values of X and A, that is, on $p(y \mid x, a)$. This 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 in a purely observational study where the observer has had no real physical control of the value of A even after the value of X was observed. This aspect can be emphasized by writing $p(y \mid \text{see}(x, a))$ for such a prediction. However, in order to arrive at a causal formulation where A would be viewed as a *cause* of Y, it is almost mandatory that one thinks of a situation where the value of A could be chosen *at will*. In the event that $A = a$, this is similarly emphasized in the notation by writing $p(y \mid \text{see}(x), \text{do}(a))$. Virtually the only conceivable way to present probabilistic evidence for the claim 'A causes Y' is by considering contrasts of the form $p(y \mid \text{see}(x), \text{do}(a)) - p(y \mid \text{see}(x), \text{do}(a'))$, where a and a' are two different values of A.

From the perspective of statistical inference, the key question is now '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. But they are not the same! In particular, the *do*-experiment does not by 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 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. One can therefore say that in the *do*-situation the joint distribution of (U, X, A, Y) is only *partly* specified.

In contrast to this, it seems both natural and necessary to assume that otherwise the probabilistic structure and description of the *see*-study can be 'lifted' to the, perhaps only hypothetical, *do*-experiment without change (cf. Lindley, 2002). In particular, one should then assume that the (prior) distribution of the unobserved variable U is the same in the hypothetical *do*-experiment as in the real *see*-study, and also that the response Y, given the covariate X, the treatment A and the unobserved background U, behaves in the same way regardless of whether the event $A = a$ was *done* or merely *seen*.

The problems therefore seem to centre around the conditional distribution $p(a|\, u, x)$, the potential fallacy for a causal analysis being that of *confounding*: in an observational study it could happen that differences between the predictions $p(y|\, \text{see}(x, a))$ and $p(y|\, \text{see}(x, a'))$ given to two individuals with different values of the A variable should actually be traced back to corresponding but unobserved differences in variable U. In an extreme situation, the 'real cause' of the different responses should then be a difference in the values of the background variable U, and different values of A would only serve as an indicator of such differences. Changing the value of the indicator by some form of *do*-manipulation would not necessarily mean that the value of the underlying variable U, and then the value of Y, would also be affected.

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 $p(y|\, x, a, u)$, actually depends on U. If the considered problem formulation does not contain such potential confounders then we can say that 'all causal model variables have been observed'. As a consequence, we would continue to consider the predictions $p(y|\, x, a)$, and contrasts between them, without paying attention to whether the value of A was obtained by manipulation/intervention, or merely by observing it.

Consider then the more interesting situation where the model actually contains an unobserved potential confounder U. 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 also be utilized in a situation in which A was thought to arise from a manipulation, that is, from a *do*-experiment. This becomes possible under the following conditional independence postulate [in essence, corresponding to the 'Back-Door Criterion' of Pearl (1995, 2000)].

Definition 1
We say that treatment assignment A is non-informative about the potential confounder variable U if A and U are conditionally independent given X, that is, $p(a \mid u, \text{see}(x)) = p(a \mid \text{see}(x))$.

This postulate has two important consequences: first, the (posterior) distribution of U based on the observations X and A does not actually depend on A, that is, we have:

Proposition 1
If A is non-informative in the sense of Definition 1, *then the two posterior distributions* $p(u \mid \text{see}(x, a))$ *and* $p(u \mid \text{see}(x))$ *are the same.*

This result justifies why it is natural to call a treatment assignment A satisfying the condition of Definition 1 non-informative. The result itself is an obvious consequence of Bayes' formula: for deriving the posterior $p(u|\, \text{see}(x, a))$ we would start from the prior $p(u)$ and consider the expression $p(x, a|\, u) = p(x|\, u)p(a|\, u, \text{see}(x)) = p(x|\, u)p(a|\, \text{see}(x))$ as the likelihood. Similarly, for deriving $p(u|\, \text{see}(x))$ we would start from the prior $p(u)$ and use $p(x|u)$ as the likelihood. But

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 $p(a|\ \mathrm{see}(x))$ appears in Bayes' formula both in the numerator and the denominator, and therefore can be cancelled.]

A consequence of the above is that if the assignment rule for A in the *see*-study satisfies the non-informativeness condition, we can replace it by another assignment rule satisfying the same non-informativeness property without changing the posterior distribution of U. In particular, we could choose the value of A at will, perhaps then following some pre-specified fixed rule that may depend on the observed covariate value X but not on the unobserved background variable U. In doing so, the posterior inference concerning U will remain the same as in the *see*-study.

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 *opt* for the imaginary *do*-experiment, corresponding to the idea that the selection of the treatment A is *optional*. In particular, this would be the case if the observed value of A had been chosen in advance. The assignment rules are then denoted by $p_{\mathrm{obs}}(a|\ u,\ x) = p(a|\ u,\ \mathrm{see}(x))$ and $p_{\mathrm{opt}}(a|\ x)$. We use the same p_{obs}-notation also for the joint distribution of $(U,\ X,\ A,\ Y)$ specified by $p(u)$, $p(x|\ u)$, $p_{\mathrm{obs}}(a|\ u,\ x)$ and $p(y|\ u,\ x,\ a)$, and the notation p_{opt} for the joint distribution where $p(u)$, $p(x|\ u)$ and $p(y|\ u,\ x,\ a)$ are retained but where the assignment rule has been changed from $p_{\mathrm{obs}}(a|\ u,\ x)$ to $p_{\mathrm{opt}}(a|\ x)$. Using this notation we can then conclude the following.

Proposition 2

If A is non-informative in the sense of Definition 1, *then*

(i) *The posterior distributions of U based on the observed data X and A are the same in both schemes, that is, $p_{\mathrm{obs}}(u|\ x,\ a) = p_{\mathrm{opt}}(u|\ x,\ a)$. Here neither of these posterior distributions depends on a.*

(ii) *The posterior distributions of U based on the observed data X, A and Y are the same in both schemes, that is, $p_{\mathrm{obs}}(u|\ x,\ a,\ y) = p_{\mathrm{opt}}(u|\ x,\ a,\ y)$.*

(iii) *The predictive distributions of Y based on observed data X and A are the same in both schemes, that is, $p_{\mathrm{obs}}(y|\ x,\ a) = p_{\mathrm{opt}}(y|\ x,\ a)$.*

This first claim is a restatement of Proposition 1. The second claim follows directly from the first, by an application of Bayes' formula, when interpreting $p_{\mathrm{obs}}(u|\ x,\ a)$ and $p_{\mathrm{opt}}(u|\ x,\ a)$ as prior distributions and then using the assumption that the conditional likelihood $p(y|\ u,\ x,\ a)$ of y is common to both schemes. The third conclusion follows at once from the first by integrating $p(y|\ u,\ x,\ a)$ as a function of u with respect to the two posterior distributions in (i) above.

Stated briefly, the first two results say that under the non-informativeness condition the posterior inferences concerning U 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 assumption, we can perform Bayesian estimation of the unknown variable U in the context of an observational study 'as if' the treatment A had been chosen in a designed experiment, according to some pre-specified rule.

This conclusion extends readily to the case where such inferences are based on data on several individuals. To consider a simple case, under a natural exchangeability assumption we can study n individuals indexed by $i = 1, 2, \ldots, n$, all described by the same probability model $p(u, x, a, y)$ that was introduced above. The data consist then of n triples of the form $(X_i,\ A_i,\ Y_i)$. In problems of this type it seems natural to think that there is some interest parameter θ that is common to all these individuals and with which the inferential problem is

primarily concerned, but that there can be, in addition, individual characteristics V_i which are relevant in the context and which, for example, could influence the corresponding treatment assignments. With this in mind we write $U_i = (\theta, V_i)$. If we now make the assumption that all treatment assignments A_i are non-informative, we can apply (ii) of Proposition 2 inductively for all n individuals, always taking the previous posterior to be the prior for the next. Finally marginalizing in the posterior, we see that $p_{\mathrm{obs}}(\theta|(x_i, a_i, y_i), i = 1, 2, \ldots, n) = p_{\mathrm{opt}}(\theta|(x_i, a_i, y_i), i = 1, 2, \ldots, n)$.

Let us now consider the statement (iii) of Proposition 2. Suppose that we, after having carried out data analysis of the above kind, would like to express our conclusions by using 'causal language'. It is then natural to consider a generic individual who is considered to be exchangeable with those in the data set and then predict what will happen to such an individual, in terms of the response Y, under different treatment assignments. In view of the fact that such responses will generally depend also on the measured covariates X, we are led to consider predictions of the form $p(y|\, x, a)$. But just as above, when performing statistical estimation, we want to be assured that contrasts between such predictions, say $p(y|\, x, a) - p(y|\, x, a')$, can in fact be interpreted in a causal fashion. In other words, we want the interpretation to be as in $p(y|\, \mathrm{see}(x), \mathrm{do}(a)) - p(y|\, \mathrm{see}(x), \mathrm{do}(a'))$. But this is precisely the interpretation that we have for the predictive distributions under the non-informativeness condition.

This set of ideas can be extended in a straightforward manner to more complicated observational schemes and experimental designs, involving (sequences of) both observed covariates and optional assignments to alternative treatments. Such designs are discussed below.

3. Causality and time: a marked point process framework

Empirical evidence derived from a statistical analysis of data when used to support a causal claim ultimately relies on comparing predictions concerning the responses Y that are issued under different real or hypothetical circumstances. Such predictions can only be made relative to the chosen (finite) set of explanatory variables or processes, which can then contain both measured and unmeasured variables realized in time before the value of Y is finally determined. With this in mind we use the term *causal field* for the collection of variables and processes, both observed and unobserved, that is going to be considered in the causal problem (cf. Mackie, 1965). Changing the set of variables that are considered will generally change the predictions and hence the contents of the contemplated causal statement. A concept similar to the causal field is the *small world* found in Savage (1972, Chapter 5).

It can be said to be axiomatic to any notion of causality that it can act only forwards in time, that is the cause must precede the effect. Given this, it is quite striking that most probabilistic/statistical formulations of causal dependence suppress time completely. But this simple requirement of time ordering is not the only aspect about causality where time is important. For example, a therapy may itself be risky, and therefore its effect on a patient may well be negative if only a short follow-up time is considered. But when the follow-up time becomes longer, the direction of the effect may be reversed as the benefits of the therapy become clear. Further issues relating to time are that it may matter *when* the therapy was given, and also *what had happened in the past*, before it was given.These aspects relating to time can be conveniently dealt with by employing the general modelling framework of *event history analysis*, see e.g. Blossfeld *et al.* (1989), Arjas & Eerola (1993), Eerola (1994) and Parner & Arjas (1999).

Considering first an individual indexed by i, we assume that a random number $n(i)$ of events that are relevant from the perspective of the causal problem have occurred to this individual at

times $0 = T_{i0} < T_{i1} < T_{i2} < \cdots < T_{in(i)}$ and have been registered in the data. At each event time T_{ik}, a vector of covariates X_{ik} is measured and a treatment A_{ik} follows immediately upon this. Hence, complete follow-up data on individual i consist of event times T_{ik} and corresponding marks or labels $Z_{ik} = (X_{ik}, A_{ik})$, $k = 1, \ldots, n(i)$, and finally of a response Y_i which, however, may not be measured on every individual. The reason for this separation in the event description is the same as above: while X_{ik} is thought to represent measured characteristics of individual i, A_{ik} signifies the result of a decision which, at least hypothetically, is optional and can be chosen according to 'free will'. The variables A_{ik} are therefore the link to counterfactual reasoning: keeping the past history fixed at its observed value, we are thinking of how different actions A_{ik} 'now' might influence the value of a future response Y_i. In principle, Y_i can be any random variable included in the model, such as some conveniently chosen function of the entire realized event sequence.

We can then think of monitoring the individual event history progressively in time, having registered by time t those event times T_{ik} that have so far already occurred, that is, $T_{ik} \leq t$, together with the corresponding marks $Z_{ik} = (X_{ik}, A_{ik})$. The cumulative individual pre-t-history can be conveniently denoted by $\boldsymbol{H}_{it} = \{(T_{ik}, Z_{ik}): T_{ik} \leq t\}$. As a convention, here we do not distinguish in the notation between such histories and the sigma fields that they generate.

An advantage of inferential approaches based on the full likelihood is that if we consider a sample of individuals and their observed event histories, then postulate that there is 'no interference between subjects' (Cox, 1958), the inference can be carried out sequentially 'one individual at a time'. Always, upon considering a new individual in the data, we can think of the follow-up as starting from a new time origin at time 0, having then stored the information obtained from the previous analyses into a corresponding (posterior) distribution of population and individual level unobservables. Using standard frequentist terminology, population or structural level variables could be called *parameters* and individual level variables *random effects*. At the present level of generality it is unnecessary to make a technical distinction between these classes, as both will in Bayesian inference be treated in a similar fashion probabilistically as random variables.

At a more technical level, we assume that the observed random variables arising from such 'earlier' analyses have been stored in a sigma field that we now denote by $\boldsymbol{H}_0$. Conditioning on $\boldsymbol{H}_0$ therefore always corresponds to conditioning on previously observed data. Similarly, the corresponding unobserved variables are thought to have generated a non-trivial sigma field denoted by $\boldsymbol{H}_0$. The sigma field generated by all 'past' variables in the model, both observed and unobserved, is denoted by $\boldsymbol{F}_0 = \boldsymbol{G}_0 \vee \boldsymbol{H}_0$. The (posterior) probability distribution on $\boldsymbol{G}_0$ obtained from the previous analyses will then serve as a prior when considering the 'next' individual, while the conditioning on $\boldsymbol{H}_0$, based on observed data, can be carried straight through.

Considering then a 'next' generic individual i, we denote, for simplicity dropping the index i from the notation, by $\boldsymbol{H}_t$ the (smallest) sigma field that contains both $\boldsymbol{H}_0$ and $\boldsymbol{H}_{it}$. Similarly, we write $\boldsymbol{G}_t$ for the (smallest) sigma field, which contains both $\boldsymbol{G}_0$ and the possible 'new' random effects that are introduced into the analysis for considering this next individual. Combining both sources of information, we write $\boldsymbol{F}_t = \boldsymbol{G}_t \vee \boldsymbol{H}_t$ for the smallest sigma field containing both $\boldsymbol{G}_t$ and $\boldsymbol{H}_t$.

Remark 1. This 'one individual at a time' convention saves us from introducing a rather elaborate vector notation for the marks used for describing simultaneously observations relating to several individuals. There is a minor drawback in that then we cannot, with this

notation, cover randomized studies in which, for example, at certain times some pre-specified proportions of individuals in the study are assigned to the treatment and placebo arms. It would not be conceptually difficult, although it would be technically somewhat involved, to carry out such a modification in our marked point process formulation. Moreover, here we have made the convention that every treatment assignment A_k is preceded by a covariate measurement X_k. This may not always be the case in practice. For example, in a randomized trial there may not be a covariate measurement every time before an intervention is made. However, in some other sampling schemes a number of covariate measurements are taken before the first choice of treatment. Such situations can be covered by slightly extending the 'mark space' of the marked point process framework, as shown in Parner & Arjas (1999). In order to keep the presentation as simple as possible we do not consider these more general situations here, however.

4. Non-informative treatment assignments and confounding

As discussed earlier, a fundamental requirement in observational studies directed towards causal conclusions is that observed association between a postulated cause and the response is not in fact due to some unobserved confounder. In order to discuss these issues in the above general setting, we first need to extend our earlier Definition of the concepts of potential confounder and non-informative treatment assignments.

Definition 2

Consider an unobserved process U_t adapted to $(\boldsymbol{G}_t)$, that is, U_t is measurable with respect to $\boldsymbol{G}_t$ for all t, and denote $\boldsymbol{U}_t = \{U_s;\, s \le t\}$.

(i) *We call $\boldsymbol{U}_t$ a potential confounder (of Y relative to $(\boldsymbol{H}_t)$) at time t if the two conditional distributions $p(y|\, \boldsymbol{H}_t)$ and $p(y|\, \boldsymbol{H}_t, \boldsymbol{U}_t)$ are not the same, that is, Y is not conditionally independent of $\boldsymbol{U}_t$ given $\boldsymbol{H}_t$.*

(ii) *We say that the treatment assignments (A_n) are non-informative (with respect to the considered causal model) if for every n the conditional distribution of A_n, given $\boldsymbol{H}_{T_n-}$ and X_n, does not depend on potential confounders $\boldsymbol{U}_{T_n-}$ before time T_n.*

A more general, and also mathematically more exact, formulation of this concept is given in Parner & Arjas (1999). It bears a close correspondence, both technically and conceptually, to the concept of non-informative (or non-innovative) censoring given in Arjas & Haara (1984), see also Andersen *et al.* (1993). Although different in form, the justification of this postulate can be said to be the same as that of the *no unmeasured confounders* assumption of Robins (1986).

Our earlier statement in Proposition 1 can now be generalized as follows: if an assignment A_n is non-informative in the above sense, then our inferences about the *past* history $\{U_s: s < T_n\}$ when based on conditioning on the observed pre-T_n-history and including the covariate X_n, will be the same regardless of whether we also condition on the treatment A_n. This really justifies why A_n is called non-informative. It should be emphasized, however, that such a statement concerns the *past* history $\{U_s: s < T_n\}$ and not the whole process (U_t). Indeed, it is perfectly legitimate that a treatment A_n will influence, even 'causally', the *future* development of some unobserved process (U_t) in the causal field, as well as the later development $\{X_k: k > n\}$ of observed covariates and, ultimately, the response Y. Thus we could perhaps say that the influence of non-informative treatment assignments A_n can be *causal* but not *inferential*. Note that for covariates X_n we have not assumed such asymmetry; a new

covariate reading can have both inferential value backwards in time and causal influence forwards in time.

As earlier, we use the notations p_{obs} and p_{opt} for the two observation schemes that are compared. They refer to models that are otherwise identical [that is, the 'generators' of (T_n, X_n), $(\boldsymbol{G}_t)$ and Y relative to $(\boldsymbol{F}_t)$ are the same], except that the treatments (A_n) are assigned differently: the obs-model corresponds to the original observational study that was assumed to have generated the data, while in the opt-scheme the treatments are assigned by an arbitrary but fixed rule where each assignment can depend on the observed past. In particular, the rule might have specified all assignments in advance. We then have the following result corresponding to the first part of Proposition 2.

Proposition 3

If the treatment assignments (A_n) *in the obs-scheme are non-informative in the sense of Definition* 2, *then the posterior inferences based on the observed data remain the same if we consider them in the opt-scheme*, *that is*, *for all* $t \geq 0$, $p_{\text{obs}}(\boldsymbol{U}_t \mid \boldsymbol{H}_t) = p_{\text{opt}}(\boldsymbol{U}_t \mid \boldsymbol{H}_t)$.

Intuitively speaking, this Proposition says that at any time t, when considering the posterior distribution of the past history of a potential confounder process, and supposing that the considered observation scheme is such that (A_n) is non-informative in the above sense, knowledge regarding the past unobserved variables in $\boldsymbol{G}_t$ remains unchanged if the assignments (A_n) are thought to have been optional. In other words, as long as (A_n) is non-informative its precise form is irrelevant for inferring $\boldsymbol{U}_t$.

The proof is most easily carried out 'one marked point at a time', that is, by first considering (T_1, X_1), A_1 and $\boldsymbol{U}_{T_1-}$ in place of X, A, and U in Proposition 2 (i). The same reasoning can be applied by induction to the largest value of n such that $T_n \leq t$, and the last step of the proof concerns the updating of the posterior on the basis of interval (T_n, t). But there the obs- and opt- schemes are trivially identical as we made above the simplifying assumption that a new treatment A_{n+1} can only be assigned with positive probability at T_{n+1}, and now $T_{n+1} > t$.

As emphasized earlier, suggesting candidates for potential confounders is limited to the causal field under consideration, and therefore to unobserved processes that the investigator/observer believes may have an effect on the response Y. Likewise, assessment of whether such processes could be considered to be non-informative in the sense of Definition 2 is a subject matter issue which cannot be resolved on the basis of statistical analysis of data.

For a given context, introducing more potential confounders while at the same time keeping the amount of observed information fixed will imply more possible choices of potential confounders and thus make it less likely that they are all non-informative in the sense of Definition 2. However, as potential confounding concerns only variables and processes that have not been observed, controlling such variables by direct observation makes it less likely that the causal analysis is confounded.

Remark 2. In Proposition 3 above, as well as in the earlier statements that we have made, the formulations have followed the Bayesian paradigm to statistical inference and referred to posterior distributions of unobserved variables or processes. This is not the only option, of course. A more traditional approach would refer to 'fixed' population or structural parameters that would then be estimated, for example, by maximum likelihood, and possibly to random effects describing unobserved individual characteristics and handled in terms of probability distributions. For the former, the key observation is that if the population parameters were seen in the role of potential confounders (U_t), then being $\boldsymbol{G}_0$-measurable, and if a

corresponding non-informativeness condition were assumed, then the likelihood contributions coming from the assignment of the treatments (A_n) would not depend on such population parameters. As a consequence, these contributions could be treated in the likelihood as proportionality factors that would not depend on the parameters, and so the maximum likelihood estimates would also not be affected by such an assignment rule. Note also that for marked point processes, when considered in the martingale/stochastic intensity framework, the likelihood always takes the same canonical 'Poisson likelihood' form. However, for the random effects, the above Bayesian formulation will apply without change.

5. Causal statements and comparison of predictions

Looking now at the contents of Proposition 3 one may wonder what bearing it has for concrete causality problems. 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, then following the same basic ideas that were, in a simple case, presented at the end of Section 2.

Suppose that we have been collecting data up to a time point τ. The observed data are then $\boldsymbol{H}_\tau$, and the corresponding unobservable variables $\boldsymbol{U}_\tau$ have generated a sigma field $\boldsymbol{G}_\tau$. According to Proposition 3 we have that if the treatment assignments (A_n) were non-informative then the posterior inferences drawn from the data, expressed in the form of $p_{\mathrm{obs}}(\boldsymbol{U}_\tau \mid \boldsymbol{H}_\tau)$, can be interpreted as if the treatments had been optional. In other words, these inferences as such are valid information about how individuals respond to different treatments in different circumstances. However, this information is not sufficient for considering causal statements before it is connected with rules according to which the treatments (A_n) are chosen.

In order to do that we now consider a generic individual that was not included in the data and on whom we could, at least hypothetically in our thoughts, enforce different treatments (A_n). For such a purpose we assume that there is a rule (also called regime) according to which the treatments are assigned. In the simplest case a rule would consist of a pre-specified list $\boldsymbol{a} = (a_1, a_2, \ldots, a_k)$ of treatments that are enforced regardless of what the individual in question has experienced earlier. More generally, each treatment can be allowed to be a function of the observed history of that individual, consisting of event times, covariate readings and earlier treatments. More generally still, a rule could be 'randomized', as long as it remains non-informative in the sense we have considered above. Denoting such a rule by $\boldsymbol{A}$, we denote the predictive distribution of the response Y, based on the observed data (on 'other' individuals), by $p_{\boldsymbol{A}}(Y \mid \boldsymbol{H}_\tau)$.

Note that before obtaining a response value Y this generic individual is thought to experience a realization of the marked point process, that is, there will be a sequence of event times T_n, covariate readings X_n and treatments A_n (the latter taken according to the chosen rule $\boldsymbol{A}$). Apart from possible fixed attributes (such as gender), which have been initially chosen to characterize the considered generic individual, these 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 over such possibly time-dependent random variables.

It may be useful to distinguish between three ingredients that jointly give rise to the predictive distributions $p_{\boldsymbol{A}}(Y \mid \boldsymbol{H}_\tau)$. The first is that, in order to have a description of how individuals behave and respond to treatments, we have to set up a probability model for the event times T_n, covariate readings X_n and response Y. As discussed in Remark 2, such a model will

generally involve population level or structural parameters whose values are unknown. The chosen assignment rule forms the second ingredient. Indeed, even when the parameter values are fixed and we can condition the probabilities on their known values, the model is incomplete as a probability description of the full marked point process $(T_n, (X_n, A_n))$ until the assignment rule $\boldsymbol{A}$ is specified. After such a rule has been chosen and the model has in this way been completed, we have a way of determining, by integration over all sample paths, the predictive distribution of Y conditionally on any fixed parameter values. Using our earlier convention that the population parameters are contained in $\boldsymbol{G}_\tau$ (in fact, already in $\boldsymbol{G}_0$) we denote this predictive distribution by $p_{\boldsymbol{A}}(Y \mid \boldsymbol{G}_\tau, \boldsymbol{H}_\tau)$. Note that in practice the information $\boldsymbol{H}_\tau$ in the data will often be redundant for considering the response Y of a generic individual if the parameter values are provided by $\boldsymbol{G}_\tau$, and then in fact $p_{\boldsymbol{A}}(Y \mid \boldsymbol{G}_\tau, \boldsymbol{H}_\tau) = p_{\boldsymbol{A}}(Y \mid \boldsymbol{G}_\tau)$. Finally, a third ingredient is needed because in reality the parameters are unknown. This is where the results obtained by statistical inference are needed, and they are here summarized in the form of the posterior distribution $p_{\text{obs}}(\boldsymbol{U}_\tau \mid \boldsymbol{H}_\tau)(=p_{\text{opt}}(\boldsymbol{U}_\tau \mid \boldsymbol{H}_\tau))$. Therefore, we compute the predictive distribution $p_{\boldsymbol{A}}(Y \mid \boldsymbol{H}_\tau)$ as the expected value $E(p_{\boldsymbol{A}}(Y \mid \boldsymbol{G}_\tau, \boldsymbol{H}_\tau) \boldsymbol{H}_\tau)$, where the expectation is with respect to the posterior $p_{\text{obs}}(\boldsymbol{U}_\tau \mid \boldsymbol{H}_\tau)$.

Having now described how to deal with an assignment rule $\boldsymbol{A}$, we can consider any two such rules of interest, say $\boldsymbol{A}_1$ and $\boldsymbol{A}_2$, and compare the corresponding predictive distributions $p_{\boldsymbol{A}_1}(Y \mid \boldsymbol{H}_\tau)$ and $p_{\boldsymbol{A}_2}(Y \mid \boldsymbol{H}_\tau)$ with each other. In practice, the necessary numerical integration can be carried out efficiently and fast by Monte Carlo simulation.

A comparison between two assignment rules can be made even more concrete by comparing simulated values drawn from these two predictive distributions, using the same random seed for both. If the response Y is real valued, then we can use for the simulations the so-called 'standard construction', which is based on the inverse of the predictive cumulative distribution function evaluated at a point drawn from the Uniform(0, 1) distribution. In particular, if the predictive distributions indicate a preference of $\boldsymbol{A}_1$ to $\boldsymbol{A}_2$ in the sense of stochastic ordering between $p_{\boldsymbol{A}_1}(Y \mid \boldsymbol{H}_\tau)$ and $p_{\boldsymbol{A}_2}(Y \mid \boldsymbol{H}_\tau)$, then such a construction will automatically lead to a pointwise comparison between simulated values where the value under $\boldsymbol{A}_1$ is at least as large as the value under $\boldsymbol{A}_2$. This is somewhat analogous to considering, in the Rubin framework, the difference between the corresponding two potential outcomes for the same individual. Such differences are not estimable from data, and indeed, due to the 'Fundamental Problem of Causal Inference' referred to earlier, one could say that they do not exist jointly in reality. Simulating values from the predictive distributions by using the same random seeds is of course also only a model-based construct, which however can provide an attractive way for comparing the consequences of applying the two rules.

Note, finally, that here we are not making use of counterfactual formulations or of corresponding random variables. This seems to distinguish us from a large body of the causality literature. However, in our view the approach proposed here should be sufficient for considering epidemiological questions because they are generic in nature. The same approach can in principle also be applied when a clinician has to decide what treatment should be given to a patient on the basis of known anamnesis and a clinical examination. Apart from possible randomization, two patients with identical backgrounds and current status descriptions will be treated in the same way as long as the protocol is not changed. But our approach cannot be applied for reasoning backwards in individual cases, for example, for deciding whether an observed outcome was caused by a given treatment, or how much longer or shorter would the life of an individual patient who did not receive a particular treatment, and is now dead, have been if that treatment had actually been given. In fact, we have some doubt that such effects can ever be evaluated in a scientifically meaningful way, whatever the data and the statistical approach.

6. Examples

In order to show how the general framework above fits to substantive questions concerning the possible existence of a causal mechanism, we now consider three 'real life' examples, each illustrating a different aspect of such problems. Our discussion below uses – as best we can – standard ways of scientific reasoning, emphasizing questions of interpretation and potential fallacies in the conclusions drawn, but without carrying out the actual model specifications in detail, or the corresponding data analysis.

Example 1
(*Does psychological stress increase susceptibility to the common cold?*)

There are several reasons for considering psychological stress as a possible cause for the common cold. According to Takkouche *et al.* (2001) at least the following causal pathways seem possible: stress can alter immune function by producing changes in the concentration of cytokines, which are molecules that mediate the response to infection. Stress can also influence health behaviour, as persons under stress may take on negative habits that place them under increased risk of infection. Finally, stress may influence a person's perception of the status of his or her own health, leading to greater sensitivity in reporting symptoms.

The question of a possible causal dependence between psychological stress and susceptibility to the common cold has even been studied experimentally. Cohen *et al.* (1991) made a prospective study in which a number of subjects were given nasal drops containing viruses that were intended to resemble doses that are common in person-to-person transmission. Here, however, we consider briefly the observational study of Takkouche *et al.* (2001) in which 1149 subjects, all belonging to the personnel of a Spanish university, were followed for stress levels and the common cold for a period of up to 1 year. At the beginning of the follow-up each participant completed an anonymous questionnaire reporting on their current stress level, common cold symptoms, and other lifestyle variables that could be viewed as being potential confounders in a study of a causal dependence between stress and the common cold. The article does not report in complete detail about what questions were asked for this purpose, but at least gender, age, professional activity (faculty or staff), smoking status, alcohol, vitamin C, and zinc intake were reported, as well as their history of allergic rhinitis and history of respiratory illness other than allergic rhinitis. For good reason, apparently, the subjects were also asked three questions relating to contacts with children: total number of children, number of children <2 years of age, and number of children who go to kindergarten. Four dimensions of stress were investigated: stressful events, negative affect, positive affect and perceived stress. Follow-up information was collected by questionnaires sent out every 10 weeks in order to identify new cases of the common cold and to update the information on stress. Follow-up was terminated at the time of the first reported episode of the common cold so that the responses could be regarded as right censored survival times.

In this case all covariate information was collected by the initial questionnaire at the beginning of the follow-up. The values of these variables can probably be assumed to remain approximately constant at their measured baseline level X_0 over the entire 1-year observation period. Different statistical models that had been adjusted according to the observed potential confounders were then considered by Takkouche *et al.* (2001), but none of them seemed important according to the criteria that were used. The key question is then whether some confounders might have gone unnoticed, that is, whether there would be some variables or processes that had not been registered and that might contribute to both the stress level and to viral infections often leading to common cold symptoms. If the questions relating to contacts with small children had been missing, one could argue that at least some part of the effect that

was attributed to psychological stress was in fact confounded. Another, although probably only remote, possibility is that, by employing different statistical models to the inference and different criteria for selecting variables into such models, the causal conclusions that were drawn in this study could have been changed.

Provided that we are willing to accept the somewhat artificial idea that stress levels can be viewed as 'non-informative treatment assignments' in the sense of Definition 2, the preceding development suggests that for an assessment of the causal effect of stress levels on the common cold we should compare predictive distributions for stress levels arising from different assignment rules. In the absence of meaningful rules, and as in this scheme there are no covariates other than those measured at the baseline, it would seem most natural to consider and compare hypothetical 'stress level regimes' that the analyst had thought of, and fixed in order to make a comparison. Then the first occurrence of the common cold could be predicted on the basis of such an assumed stress level sequence in the past.

A referee suggested that we should also consider the possible applicability of our marked point process framework in situations in which there can be an 'intermediate marker' event, such as measurement of a surrogate variable, between the treatment and the actual endpoint. (Such surrogate markers can be important in epidemiological studies; for example, in a cohort study the actual endpoint could be coronary heart disease or stroke, but in the absence of those, because of right censoring, one could consider using blood cholesterol level as a surrogate endpoint.) Here the referee suggested that the concentration of cytokines mentioned at the beginning of this example could be considered as an intermediate marker, and wrote, 'It would be of interest to enquire if the mechanism by which stress affects the common cold is through the immune response and/or through other pathways.'

In order to fix ideas, suppose – again purely hypothetically – that the concentration of cytokines was monitored in the context relevant method by carrying out corresponding measurements at the same 10-week intervals as the information on stress levels and on new cases of the common cold was updated. It is then natural to think of these measured concentrations in the marked point process model as covariates X_n. Indeed, as the values of X_n are themselves random and cannot therefore in reality be known or chosen in advance, it would seem difficult think about them in the role of treatment assignments. Even less credible, on this very crude level of observed information, would be the idea that they would be non-informative in the sense of Definition 2. Closer knowledge of the underlying biochemical pathways/processes, and their monitoring as covariates, might naturally change the situation.

In order to respond to the referee's request, suppose then that we have decided to treat stress levels as treatment assignments A_n and cytokine concentrations as covariates X_n. By applying sequentially the well-known methods of conditioning on past events we can now model the development of the marked point process $(T_n, (X_n, A_n))$, and of the corresponding response Y. Having estimated such probabilities from data we can consider, and compare for the purpose of assessing the role of the covariates (X_n), two kinds of predictions of Y at time points T_n: first, predictions which are based only on the baseline covariate information X_0 and on the sequence of past stress levels A_n, $k \leq n$ as described above, and second, predictions using information that in addition contains the past measured cytokine concentrations X_k, $k \leq n$. A comparison of these two predictions would give us some idea about how useful it would be to know, in addition to the stress levels, the corresponding levels of cytokine concentration. If the predictions would be much improved by such knowledge, this would be evidence that the cytokine concentrations would be closely associated with processes belonging to 'the causal pathway from stress to the common cold'. Moreover, if their inclusion in the predictions would make knowledge of stress levels practically redundant (which is a conditional independence statement), then this would provide an even stronger motivation for studying the relationship of cytokines to this pathway.

However, based only on such statistical considerations, and without more substantive knowledge of the relevant biochemical processes, we cannot rule out the possibility that the measured cytokine concentrations could still be only markers influenced by some more basic latent processes belonging to the causal pathway.

Example 2
(*Does pregnancy cause cohabiting couples to get married?*)

As a second illustration, consider the question of whether, for cohabiting couples that have not been previously married, pregnancy of the woman has a causal influence on their 'time to marriage' (see e.g. Blossfeld & Rohwer, 1995; Blossfeld *et al*., 1999; Arjas, 2001). For a statistical analysis of this question, there might be demographic data on potentially relevant 'fixed' covariates, such as the ages of the cohabiting partners, their social class, level of education, etc. Follow-up data from the beginning of the cohabitation status would then register the time to pregnancy, if any, and the time to marriage or alternatively to the end of cohabitation or the end of follow-up.

In order to study this question, one obviously first has to think about the influence of the measured covariates on the couple's 'rate to get married'. Suppose that there were enough of such demographic data that one could either stratify according to these covariates or construct a sensible statistical model to account for their influence. In either case, it would be natural for an outsider to consider couples to be *a priori exchangeable* if they share the same measured covariate values. Based on such data one could then estimate 'marriage rates' for groups of couples with shared covariate values. In particular, one can compare such rates before and after pregnancy was established, as functions of the time that they had lived together. Empirical results of Blossfeld *et al*. (1999) show that the marriage rate starts to increase soon after pregnancy has been established, thereby also changing the predictive distribution of 'time to marriage'. After some time the marriage rate again starts to decrease and, at the time the child is born, it is already close to the marriage rate of couples with similar covariate history except that they had not conceived a child.

Suppose for simplicity that there is so much data that a frequentist and a Bayesian analysis would have produced essentially the same marriage rate curves. From a frequentist point of view, their values would be interpreted as marriage frequencies averaged over a population, and from a subjectivist perspective they could be said to consist of momentary predictions concerning the marriage of 'a generic couple of unknown intentions' belonging to the exchangeable group of couples that is specified in terms of their shared covariate values. The Definition of such a group would obviously depend on what covariate information would be provided in the data.

The question is now whether we can draw some interesting and valid causal conclusions from such statistical results. A straightforward claim in this situation would be to say that the observed differences in the marriage rates before and after established pregnancy were actually evidence that 'pregnancies are causing marriages'. Such a conclusion does not seem to be completely warranted, however, because there are likely to be important unobserved differences between couples that then have the potential of confounding such a causal analysis. Presently several effective and affordable contraceptive methods are available, and therefore unplanned pregnancies for couples living in a stable relationship are likely to be quite rare. However, most couples in a fertile age that want a child will be able to conceive one without having to wait for very long. Therefore, most likely, both pregnancy and marriage are premeditated events for which plans have existed in the minds of the cohabiting partners well before either conception or marriage.

On the level of individual couples, considering pregnancy as a cause of marriage would necessitate comparing the two options 'pregnancy' and 'no pregnancy', but keeping the identity of the couple, including their plans and intentions, fixed. Conceptually, think of a situation in which the contraceptive method that was used had failed, resulting in an unplanned pregnancy, and in which abortion would not be an option could perhaps approach such a setting. Such a comparison, however, is impossible in a study based only on demographic data because the plans and intentions of the couples remain unknown, and therefore they become potential confounders of the causal analysis. A completely plausible explanation of the observed data, without any true causal mechanism, is that the estimated marriage rates before and after pregnancy are merely a reflection of a selection process taking place over time: those who intended to marry did so either already before pregnancy was established or soon after. Established pregnancy would therefore *not* be non-informative about such plans, but rather serve as an indication that the couple were intending to stay together and perhaps marry. In the estimation of the marriage rates such couples were likely to be progressively 'selected away from the risk set', whereas those who did not intend to marry also did not do so after established pregnancy. This explains the downward trend in the observed marriage rate.

This, of course, does not say that, in some cases, pregnancy could not come as a surprise and then act as the direct cause of a decision to marry. In former times, such a sequence of events must have been quite frequent. Today, however, a causal claim would seem better justified if one were considering, for example, the issue of whether remaining childless could be considered as a cause of divorce. The basis of such a claim could be a similar comparison as above, but now between divorce rates before and after the birth of the first child (assuming that the couple did not have children already when they married). From a causal perspective, the important difference between this and the previous question is in that here it would not be unnatural to assume that a large majority of all couples entering marriage would want to have children. This hypothesis (which of course cannot be checked from the data either), if true, would effectively rule out the possibility that the contemplated cause, birth of a child, or lack of it, was in fact to some degree already determined by an earlier unobserved confounder variable.

7. Concluding remarks

The main point in the above development has been an attempt to provide a way in which causal questions and statistical inference could be considered within a single probabilistic framework. Compared with the extensive literature discussing the role of statistics in causal reasoning, there appears both to be several points in common as well as points of departure. This seems to be typical, perhaps due to the strong philosophical element of causality problems, and often papers differ from each other on the level of foundational issues. Perhaps therefore the opinions expressed are often in sharp conflict with each other, with little willingness on the part of the authors to accept alternative problem formulations and corresponding solutions (see, e.g. the discussion of Dawid, 2000).

As in the work of Pearl (1995, 2000) and Dawid (2000, 2002), our approach has been entirely probabilistic and the interpretation of probability is openly subjective. A further common aspect is the emphasis on predictive probabilities/distributions as the most important criterion for assessing causal effects. The main distinction is that we have not used graphical modelling as a tool. A further difference from Pearl, as well as from many works of Robins, is that we have not made use of structural modelling. Furthermore, unlike Rubin and Robins, and many other authors who have followed their path, we have not applied the idea of potential outcomes and introduced corresponding random variables. In the case of sequential

treatment regimes, putting aside the fact that our inferential approach differs from those employed by Robins and that we perform an integration with respect to the posterior, our use of predictive distributions corresponds to Robins' so-called G-computation formula in the simple case in which the regimes are deterministic. In more general situations, while we compare the consequences of applying two such rules without specifying what particular treatments will then be used, Robins makes a comparison between specific treatment sequences. In this sense, one could perhaps say that the G-computation formula makes a post-intervention assessment considering 'treatments actually received'. For a recent contribution to the area of sequential assignments, see Didelez (2003).

Adopting an openly subjectivist view of probability obviously gives a great deal of freedom, as numerical values given to probabilities are no longer viewed as objective claims about Nature, including humans, but as statements of what 'I', as an investigator, think. A critic of the approach might then say that by such a method nothing can be ever proven, and that Bayesian statistics is only an organized way of combining data and opinions for the purpose of forming further opinions. In particular, causal dependencies cannot be shown to be true by such methods. This may be so and, indeed, we do not believe that causal relationships can be 'proven' by adopting sufficiently clever statistical techniques or procedures (which includes those presented above). But it would be unfair to say that such a subjectivist position implies that one is making causal claims that are arbitrary. Actually, the opposite is true in the sense that then the methods of statistical inference, being completely based on probability calculus, have a strong normative element.

A second point is that issues involving subjective judgement are unavoidable in any case: whatever the statistical approach, there has to be some condition that either rules out the possibility of confounded explanation of the observed responses, or at least restricts the influence of potential confounders in estimating the causal effects. Such conditions cannot be verified only by analysing data. Pearl actually makes a further distinction between probabilistic and causal judgements, saying that judgements that lead a person to consider a particular causal field, and then postulate conditions like those appearing in our Definition 2, are causal in nature, not probabilistic. Be it as it may, we believe that the required judgements concerning potential confounding can be made reasonably only in a considered substantive context, and that they influence the final conclusions regarding causal relationships far more than, say, the choice, within reasonable limits, of Bayesian prior distributions for the structural parameters. It is then a big bonus if such judgements can be related to a model that has been formulated in way that is intuitively meaningful and understandable even outside a small circle of statistical experts.

Acknowledgements

We are grateful to Niels Keiding, Judea Pearl, Jamie Robins and Don Rubin for discussions on some earlier versions of this paper, and to two referees for their suggestions for a revision.

References

Andersen, P. K., Borgan, O., Gill, R. D. & Keiding, N. (1993). *Statistical models based on counting processes*. Springer, New York.

Arjas, E. (2001). Causal analysis and statistics: a social sciences perspective. *Eur. Sociol. Rev.* **17**, 59–64.

Arjas, E. & Eerola, M. (1993). On predictive causality in longitudinal studies. *J. Statist. Plann. Inference* **34**, 361–386.

Arjas, E. & Haara, P. (1984). A marked point process approach to censored failure data with complicated covariates. *Scand. J. Statist.* **11**, 193–209.

Blossfeld, H.-P., Hamerle, A. & Mayer, K. U. (1989). *Event history analysis*. Erlbaum, Hillsdale, NJ.
Blossfeld, H.-P., Klijzing, E., Pohl, K. & Rohwer, G. (1999). Why do cohabiting couples marry? An example of causal event history approach to interdependent systems. *Qual. Quant.* **33**, 229–242.
Blossfeld, H.-P. & Rohwer, G. (1995). *Techniques of event history modeling. New approaches to causal analysis*. Erlbaum, Mahwah, NJ.
Cohen, S., Tyrrell, D. A. & Smith, A. P. (1991). Psychological stress and susceptibility to the common cold. *N. Engl. J. Med.* **325**, 606–612.
Cox, D. R. (1958). *Planning of experiments*. Wiley, New York.
Dawid, A. P. (2000). Causal inference without counterfactuals (with Discussion). *J. Amer. Statist. Assoc.* **95**, 407–448.
Dawid, A. P. (2002). Influence diagrams for causal modelling and inference. *Int. Statist. Rev.* **70**, 161–189.
Didelez, V. (2003) Graphical models and sequential decisions. *B. Int. Statist. Inst. 54th Session Book* **1**, 148–151.
Eerola, M. (1994). *Probabilistic causality in longitudinal studies*. Lecture Notes in Statistics. Vol. 92. Springer-Verlag, Berlin.
Freedman, D. (1999). From association into causation: some remarks on the history of statistics. *Stat. Sci.* **14**, 243–258.
Holland, P. W. (1986). Statistics and causal inference (with Discussion). *J. Amer. Statist. Assoc.* **81**, 945–970.
Hume, D. (1739). *A treatise of human nature*. John Noon, London.
Lindley, D. V. (2002). Seeing and doing: the concept of causation. *Int. Statist. Rev.* **70**, 191–214.
Mackie, J. L. (1965). Causes and conditions. *Am. Philos. Quart.* **4**, 245–264.
Neyman, J. (1923). On the application of probability theory to agricultural experiments. Essay on principles, section 9. *Statist. Sci.* **5**, 465–480 (Translated, 1990).
Parner, J. & Arjas, E. (1999). Causal reasoning from longitudinal data. Research Report A27 of Rolf Nevanlinna Institute, University of Helsinki, Helsinki.
Pearl, J. (1995). Causal diagrams for empirical research (with Discussion). *Biometrika* **82**, 669–710.
Pearl, J. (2000). *Causality*. Cambridge University Press, Cambridge.
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. Modelling* **7**, 1393–1512.
Robins, J. (1997). Causal inference from complex longitudinal data. In *Latent variable modeling and applications to causality*, Lecture Notes in Statistics, Vol. 120, 69–117. (ed M. Berkane) Springer-Verlag, New York.
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 treatments. *J. Educ. Psychol.* **66**, 688–701.
Rubin, D. B. (1978). Bayesian inference for causal effects: the role of randomization. *Ann. Statist.* **6**, 34–58.
Savage, J. (1972). *The foundations of statistics*. Dover, New York.
Spirtes, P., Glymour, C. & Scheines, R. (1993). *Causation, prediction and search*. Springer-Verlag, New York.
Takkouche, B., Regueira, C. & Gestal-Otero, J. J. (2001). A cohort study of stress and the common cold. *Epidemiology* **12**, 345–349.

Received December 2002, in final form November 2003

Elja Arjas, Rolf Nevanlinna Institute, University of Helsinki, PO Box 4, FIN-00014 Helsinki, Finland.
E-mail: elja.arjas@rni.helsinki.fi