Causality, Confounding, and Control

Michel MOUCHART [a] , Federica RUSSO [b] and Guillaume WUNSCH[c]

[a] *Institute of Statistics, Université catholique de Louvain, Belgium*

[b] *Philosophy, SECL, University of Kent at Canterbury, UK.*

[c] *Institute of Demography, Université catholique de Louvain, Belgium.*

April 26, 2006

Draft in progress

Comments welcome

1 Introduction

In a previous paper, Russo *et al.* (2006), causality is considered in the framework of structural models, *i.e.* statistical models characterized by parameters that are stable over a large class of interventions or of environmental changes and that take into account background and contextual knowledge. From this statistical viewpoint, causality is defined in terms of exogeneity in a structural model. This approach allows us to attain a concept of causality that is internal or relative to the structural model itself. Thus our knowledge of causal relations depends on structural models that mediate epistemic access to causal relations.

In the social sciences, structural models correspond to causal structures that are much more complex than simple causal relations such as X causes Y. Russo *et al.* (2006) mentions some practical difficulties such as covariate sufficiency, no confounding, *etc.* without developing a systematic analysis of those issues. In the present paper we examine more thoroughly the concept of exogeneity and the problem of confounding and control. More specifically, we show that confounding due to latent variables raises substantial problems for exogeneity. We start by examining the issue of exogeneity in a conditional structural model and then proceed with the concept of confounding and its impact on causal modelling. We conclude the paper with a brief discussion on how and when to control for confounding.

Exogeneity has a long history in the foundations of statistical modelling. In econometrics, Hood and Koopmans (1953) gathers a number of pionneering contributions from the Cowles Foundation, Engle *et al.* (1983) and Florens and Mouchart (1985), among others, elaborate further these con-

tributions. In spite of this long history no consensus on this concept seems to emerge. Spirtes *et al.* (2000), for instance, gives three formal definitions in their monograph on causality and graph models, the equivalence of which is not demonstrated. In the social sciences, a vast literature on LISREL type models, or Covariance Structure Models, makes use of the concept of exogeneity, see *e.g.* Bollen (1989, p.12): "exogenous because its causes lie outside the model", p.46: "exogenous ... means that the variable is determined outside the model". We shall argue that the concept of exogeneity requires, as a prerequisite, a proper understanding of what a conditional model is about; failing to make the meaning of "determined outside the model" explicit hinders an adequate handling of missing data and/or unobservable (or, latent) variables.

Confounding is not a new issue either. In philosophy, Reichenbach (1956) already developed the concept of *conjunctive fork* where two or more effects have a common cause and where these effects are conditionally independent given the common cause. The example of a drop in atmospheric pressure causing both a storm and a barometer dip is well-known. Simpson's paradox too is a classic example of confounding (Rouanet 1985). The presence or absence of confounders is also a major issue in epidemiology. If confounders are not taken into account, how can one infer true causal relations and develop preventive or curative health policies? In demography, one knows since the end of the 19th century that the differences in mortality between countries or provinces might be due to their differential age structure only. In this case, the latter confounds the true geographical mortality pattern. In all these cases, the recommended remedy is to control for the confounding factor, *e.g.* the atmospheric pressure in the weather example or the age structure in the demographic case. It is worth pointing out, following Pearl (2000) and Dawid (2002), that one should distinguish between conditioning by observation or by intervention. The section on control will only consider the conditioning approach on observational data, more suited to population studies where intervention or manipulation is rarely possible. For the issue of conditioning by intervention and the use of influence diagrams, see Dawid (2002).

2 Exogeneity and Confounding

2.1 Conditional Models

Originally, the concept of exogeneity appears with regression models. A first, and naive, approach was to consider an exogenous variable as a non-random variable, the endogenous variable being the only random one. That this approach was unsatisfactory became clear considering complex models where the same variable could be exogenous in an equation and endogenous in another. A first progress came through a proper recognition of the nature of a conditional model. Here, we present a heuristic account of the basic concepts; for a more formal presentation, see *e.g.* Mouchart and Oulhaj (2000) and Oulhaj and Mouchart (2003).

Let us start with *an (unconditional) parameterized* statistical Model $\mathbf{M}_{\mathbf{X}}^{\omega}$ given in the following form:

$$\mathbf{M}_{\mathbf{X}}^{\omega} \;=\; \{p_X(x \mid \omega) \;:\; \omega \in \Omega\} \tag{1}$$

where for each $\omega \in \Omega$, $p_X(x \mid \omega)$ is a (sampling) probability density on an underlying sample space corresponding to a (well-defined) random variable X and Ω is the parameter space, aimed at describing the set of sampling distributions considered to be of interest. A conditional model is constructed through embedding that concept into the usual concept of an unconditional statistical model (1). For expository purposes, this paper only considers the case where a random vector X of observations is decomposed into $X' = (Y', Z')$ (where ' denotes transposition) and the model is conditional on Z.

The basic idea of a conditional model is the following: starting from a global model $\mathbf{M}_{\mathbf{X}}^{\omega}$ as given in (1), each sampling density $p_X(x \mid \omega)$ is first decomposed through a marginal-conditional product:

$$p_X(x \mid \omega) \;=\; p_Z(z \mid \phi)\, p_{Y|Z}(y \mid z, \theta) \qquad\qquad \omega \;=\; (\phi, \theta) \tag{2}$$

where $p_Z(z \mid \phi)$ is the marginal density of Z, parametrized by ϕ, and $p_{Y|Z}(y \mid z, \theta)$ is the conditional density of $(Y \mid Z)$, parametrized by θ. Next, one makes specific assumptions on the conditional component leaving virtually unspecified the marginal component. Thus a conditional model may be represented as follows :

$$\mathbf{M}_{\mathbf{Y}}^{\mathbf{Z},\theta;\Phi} \;=\; \{p_X(x \mid \omega) \;=\; p_Z(z \mid \phi)\, p_{Y|Z}(y \mid z, \theta) \qquad\qquad \omega = (\theta, \phi) \in \Omega = \Theta \times \Phi \;\} \tag{3}$$

where Φ parametrizes a, typically large, family of sampling probabilities on Z only and for each $\theta \in \Theta$, $p_{Y|Z}(y \mid z, \theta)$ represents a conditional density of $(Y \mid Z)$. The essential features of a conditional model are therefore:

1. θ indexes a well specified family of conditional distributions. This family constitutes the kernel of the concept of a conditional model. The concept of conditional model relates, however, to a family of joint distributions $p_X(x \mid \omega)$ obtained by crossing the family of conditional densities $p_{Y|Z}(y \mid z, \theta)$ with a family of marginal distributions $p_Z(z \mid \phi)$.

2. ϕ is a nuisance parameter which is identified by definition (because Φ is a set of distributions of Z). Furthermore θ and ϕ are variation free. The notation $\mathbf{M}_{\mathbf{Y}}^{\mathbf{Z},\theta;\Phi}$ conveys the idea that θ is the only parameter of actual interest, leaving to ϕ no explicit role.

3. The modelling restrictions are concentrated on the conditional component, *i.e.* the set $\{P_Y^{Z,\theta} : \theta \in \Theta\}$ embodies the main hypotheses of the model, whereas in most cases, the set Φ embodies a minimal amount of restrictions, typically only the hypotheses necessary to guarantee essential properties for the inference on θ, such as identifiability or convergence of estimators.

Consequently, in most situations, but not in all, Φ represents a "thick" subset of the set of all probability distributions of Z. The role of Φ is to stress the random character of Z at the same time as the vague specification of its data generating process; Φ may nevertheless play an important role because its specification may determine desirable properties of the estimators of θ, the parameter of interest. (Oulhaj and Mouchart(2003) provides more information on the very nature of conditional models)

2.2 Exogeneity and conditional model

Suppose we analyze data $X = (Y, Z)$. A challenging issue is to decide whether it is admissible, in the sense of losing no relevant information, to only specify a conditional model $\mathbf{M}_{\mathbf{Y}}^{\mathbf{Z},\theta;\Phi}$ rather than specifying the model $\mathbf{M}_{\mathbf{X}}^{\omega}$. This is the issue of exogeneity.

The motivation for specifying a conditional model rather than a model on the complete data X is parsimony: some specifications on the marginal process may not be avoided for ensuring suitable properties of the inference on the parameters of the conditonal process but by specifying less stringently the marginal process, generating Z, one looks for protection against specification error. The cost could however be substantial if the marginal process contains relevant information.

Formally, the condition of exogeneity is therefore: the parameter of interest should only depend on the parameters identified by the conditional model and the parameters identified by the marginal process should be "independent" of the parameters identified by the conditional process. Here, "independence" means "variation-free" in a sampling theory framework or independent in the (prior) probability in a Bayesian framework. It should be stressed that the independence among parameters has no bearing on a (sampling) independence among the corresponding variables.

More specifically, let us consider the following marginal-conditional decomposition:

$$p_X(x \mid \omega) = p_Z(z \mid \theta_Z)\, p_{Y|Z}(y \mid z, \theta_{Y|Z}) \tag{4}$$

where θ_Z , resp. $\theta_{Y|Z}$, represents the parameter identified by the marginal , resp. conditional, process. The condition of independence, namely:

$$(\theta_Z, \theta_{Y|Z}) \in \Theta_Z \times \Theta_{Y|Z} \quad \text{or} \quad \theta_Z \perp\!\!\!\perp \theta_{Y|Z} \tag{5}$$

is a *condition of (Bayesian) cut* (see Barndorff-Nielsen (1978) in a sampling theory framework and Florens *et al.* (1990) in a Bayesian framework), and is deemed to allow for a separation between the inference on the parameters of the marginal process and the inference on the parameters of the conditional process. More explicitly, condition (5) implies that any inference on θ_Z, resp. $\theta_{Y|Z}$, be based only on the marginal, resp. conditional, model characterized by the marginal distributions $p_Z(z \mid \theta_Z)$, resp. conditional distributions $p_{Y|Z}(y \mid z, \theta_{Y|Z})$.

This condition along with the condition that the parameter of interest, say λ, depends only on the parameters identified by the conditional process , *i.e.* $\lambda = f(\theta_{Y|Z})$, formalizes the concept of

"losing no relevant information" when basing the inference on the conditional model rather than on the complete model, characterized the distributions $p_X(x \mid \omega)$. In this setting, the concept of exogeneity appears as a binary relation between a function of the data, namely Z, and a function of the parameters, namely λ. Thus, Florens *et al.* (1990) suggests the expression "Z and λ are mutually exogenous" (or Z is exogenous for λ), to stress the idea that a variable is not exogenous by itself but is exogenous in a particular inference problem. Treating Z as exogenous means therefore that the (marginal) process generating Z is minimally specified (and may be heuristically qualified as "left unspecified") *and* that the inference on the parameter of interest, although based on the joint distribution of all the variables in X, is nevertheless invariant with respect to any specific choice of the marginal distribution of Z.

Summarizing: exogeneity is the condition that makes admissible the use of the conditional model $\mathbf{M}_{\mathbf{Y}}^{\mathbf{Z},\theta,\mathbf{\Phi}}$ as a reduction of the complete model $\mathbf{M}_{\mathbf{X}}^{\omega}$; furthermore, exogeneity is a property at the level of a conditional *model* rather than at the level of a particular variable.

The consequences of a failure of exogeneity may be twofold. There may be a loss of efficiency in the inference if the failure comes from a restriction (equality or inequality), or a lack of independence in a Bayesian framework, between the parameters of the marginal model and those of the conditional model. There may be also an impossibility of finding an unbiased or a consistent estimator if the parameter of interest is not a function of $\theta_{Y|Z}$ only. A typical example, well known in the field of simultaneous equations in econometrics, is that the parameter of interest in a structural equation may not be a function of the parameters identified by the conditional model corresponding to the equation.

2.3 Exogeneity and Causality

In general, the specification of a parameter of interest is a contextual rather than a statistical issue. A most usual rationale for specifying the parameter of interest is based on the notion of a *structural model*. Following Russo *et al.* (2005), this is a model deemed to represent an underlying structure of a data generating process. More explicitly, apart from adequately fitting the data, a structural model should also be invariant under a large class of changes of the environment (or, of interventions) and should, at the same time, be in agreement with the knowledge of the field from which the data are extracted. In other words, such a model should be structurally stable and reflect the scientific knowledge of the field.

In this framework, Russo *et al.* (2005) approaches causality as *exogeneity in a structural conditional model*. In the very simple case of two variables Y and Z, this concept may be paraphrased as follows: "if the conditional distribution of Y given Z is structurally stable and reflects a good scientific knowledge of the field, there is no reason not to consider that Z causes Y". This approach might be considered an empirical one, as long as the observations providing the ground for a causal

interpretation are not only the data under immediate scrutiny but also the whole body of observations underlying the "field knowledge" and leading accordingly to the present state of scientific knowledge. In this sense, causal attribution "Z causes Y" is an issue of statistical modelling, namely this is the question whether the conditional model characterized by $p_{Y|Z}(y \mid z, \theta_{Y|Z})$ is actually *structural*.

3 Confounders and Confounding

In a recent paper on trends in cardiovascular diseases (CVD) in Europe, Kesteloot *et al.* (2006), the authors attribute the impressive current decline in mortality in the Baltic States to a change in dietary habits, *i.e.* the greater consumption of vegetable oil for cooking and the progressive replacement of butter by low-fat margarine. Though nutrition does indeed play a significant role on the incidence of CVD, one may also postulate, Gaumé and Wunsch (2003), that, in the case of the Baltic States, the dramatic change from a communist regime to a liberal one has led to systemic repercussions in the whole society which have brought about both modifications in cardiovascular mortality (through major fluctuations in stressful events), in economic conditions, and in behaviours including nutrition. Societal transformations and contextual changes in the economic, social (including public health), and political spheres would therefore be a confounder masking the true relation between changes in mortality patterns and in nutritional ones. To put it simply, correlation of time series does not imply causation.

To take another example that will be discussed later in this paper, medical reports in the 1920s already pointed out the suspected links between tobacco and cancers, and a 1938 article in the journal *Science* suggested that heavy smokers had a shorter life expectancy than nonsmokers. In 1939, F.H. Müller also published a paper in German on the relationship between smoking and lung cancer, see Bartecchi *et al.* (1995) and Freedman (1999). Though one now knows that smoking is bad for one's health, actually it is only since the early 1950s that a series of studies had established the fact, Vallin *et al.* (2006), and even then the relationship between lung cancer and cigarette smoking was hotly disputed by R.A. Fisher who also argued that correlation is not causation. When is a correlation causal and when is it the result of confounding? This is the issue tackled here, by recalling some well-known and lesser-known facts.

In the following, we will use, as a visual aid, causal directed graphs where nodes represent variables and directed edges (single-headed arrows) represent the possible impact of the variable at the base of the edge on the variable at the head of the edge. Here, "impact" refers to a conditioning variable in a structural conditional model.

In epidemiology and in demography, when one examines the impact of a treatment or exposure on a response or outcome, a confounding variable - or confounder - is often defined as a variable

associated both with the putative cause and with its effect, see *e.g.* Jenicek and Cléroux (1982), Elwood (1988). Sometimes the definition is more precise, such as in Anderson *et al.* (1980) or in Leridon and Toulemon (1997). According to these authors, a variable or background factor is a confounder whenever two conditions simultaneously hold:

1. The risk groups differ on this variable;

2. the variable itself influences the outcome.

Some authors gloss condition 1 adding that the background factor should not be a consequence of the putative cause, Schlesselman (1982).

For instance, if we examine the impact of cigarette smoking on the incidence of cancer of the respiratory system, a variable such as exposure to asbestos dust confounds the relation between smoking and this type of cancer. Exposure to asbestos dust and smoking are associated, *i.e.* there are proportionally more persons exposed to asbestos in the smoking group than in the non-smoking group. Condition 1 is therefore satisfied. In addition, inhaling asbestos dust is a strong cause of cancer of the pleura; condition 2 is thus also satisfied. Cancer is the outcome variable in this example, smoking a potential cause, and exposure to asbestos a confounder. Vice-versa if one were to examine the impact of asbestos exposure on the incidence of cancer of the respiratory system, smoking this time would be the confounding factor, as it is associated with asbestos exposure and is a cause of lung cancer. This simplified example is developed in Russo *et al.* (2006); a real study would also consider other causal factors and paths, and the interaction (or effect modification) between smoking and asbestos exposure.

Condition 1 needs to be clarified however. Why are smoking and asbestos exposure associated? Reichenbach (1956) was one of the first, if not the first, in philosophy to point out that simultaneous correlated events must have a prior common cause: "If improbable coincidence has occurred, there must exist a common cause" (p.157). At around the same time, statisticians were also aware that a correlation between two variables could be due to a common cause. Considering once again the correlation between smoking and lung cancer, suppose that one's unknown genotype (G) would influence both smoking behaviour or tabagism (T) and the susceptibility to lung cancer (C). This explanation was proposed by the statistician R.A. Fisher when he was scientific consultant to the Tobacco Manufacturers' Standing Committee in the late 1950s (Fisher 1957). In this case "without any direct causation being involved, both characteristics might be largely influenced by a common cause, in this case the individual genotype", Fisher (1958). The corresponding causal graph would be as in Fig. 1 with possibly no causal link at all between T and C.

Fisher's suggestion was ultimately rejected but it did show that proving the relation between lung cancer and cigarette smoking required not only sound epidemiological evidence, but also replication in different studies, experimental evidence from animal studies, and a plausible biological mechanism

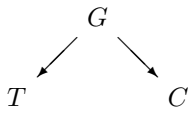

Figure 1: *Genotype, smoking and tabagism*

linking smoking to lung cancer. This is precisely what we mean by structural modeling.

Coming back to the question 'why are smoking and asbestos exposure associated?', one knows in demography and in epidemiology that both smoking and asbestos exposure are dependent upon one's socio-economic status (SES): those with a lower SES tend more to smoke and work in unhealthy environments than those with a higher SES. The causal graph can therefore be drawn as in Fig. 2, where A represents exposure to asbestos, T tabagism, and C cancer incidence. Notice that Fig. 2 incorporates two assumptions, namely: $A \perp\!\!\!\perp T \mid SES$ and $C \perp\!\!\!\perp SES \mid A, T$.

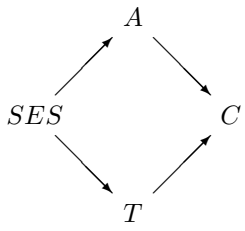

Figure 2: *Socio-economic status, smoking, absbestos exposure and cancer of the respiratory system*

This graph shows that tabagism and asbestos exposure are in fact not independent from one another as they are both related to one's SES, *i.e.* they have a common cause. Note that SES is also a common cause of T and C as it has an impact on cancer through the intervening or intermediate variable A. However an association between two variables such as smoking and asbestos exposure could also be due to a causal relation between them. A could be a cause of T or vice-versa. The two corresponding causal graphs are given in Fig. 3 and 4 respectively.

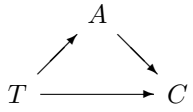

Figure 3: *The relation between T and C, A being an intervening variable*

This distinction leads to a more precise definition of a confounder: a confounding variable or confounder is a variable which is a common cause of both the putative cause and its outcome, Bollen (1989), Pearl (2000). In graphic representations, a common cause is a common ancestor to both

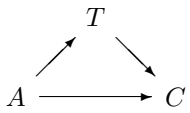

Figure 4: *The relation between T and C, A being a common cause*

putative cause and effect. For example, A is a confounder in Fig. 4 because in this model it is a common cause of both T and C. For the same reason, SES is a confounder in Fig. 2, as it is a common cause of both T and C (the latter via A). In Fig. 3, A is not a common cause of T and C; A is therefore not a confounder. The confounder can be either latent (*i.e.* unobserved) or observed: this aspect is developed in Section 5. From the point of view of controlling for the confounder, the two cases are obviously quite different, see Cox and Wermuth (2004). This definition avoids taking an intervening (intermediate) variable between the putative cause and the outcome such as in Fig. 3 as a confounder, even though it is associated with the putative cause (as the latter has a causal influence on the former) and it has an impact on the outcome. Many definitions given in epidemiology textbooks are not adequate in this respect. As one can presume, the possible confounder should not be affected by treatment/exposure, Schlesselman (1982). Finally, using the so-called d-separation or d-connection criteria, it is possible in causal graphs to check for confounding, see Pearl (2000), Robins (2001).

4 Reichenbach on screening-off

4.1 What did Reichenbach really said?

The history of the concept of screening-off traces back to Hans Reichenbach, who first coined the term. But what was Reichenbach talking about? What did he use "screen-off" for? Screening-off and common cause often go *pari passu*, but we'd better go back reading Reichenbach because this is not exactly the case. To begin with, Reichenbach develops his probabilistic theory of causality as part of his causal theory of time. Causes do not precede effect by definition, on the contrary, Reichenbach aims to develop a theory of causal relations implying causal asymmetry, which in turn will be used to define the relation of temporal priority of the causes and thus the direction of time and time ordering of events. In Chapter IV, Reichenbach extends the discussion to macrostatistics, that is "to processes the elementary 'particles' of which are macroscopic objects (such as grains of sand or playing cards) and the elementary arrangements of which are not microstates but macrostates" (Reichenbach 1959, p. 145). Applications of marcrostatistics in this chapter and the resulting principle of common cause (PCC) in §19 will eventually lead to the definition of the time direction.

Let us turn the attention to PCC. The problem to be solved is explaining improbable coinci-

dences or correlations. The improbable, according to Reichenbach, has to be explained in terms of causes, not in terms of effects (1956, §19). Thus the Principle of Common Cause: *If an improbable coincidence has occurred, the must exist a common cause* (Reichenbach 1956, p.157). More precisely, fortuitous coincidences are not impossible and the existence of a common cause is not certain either, but only probable. However, as long as those coincidences occur, the probability of the existence of a common cause raises. On top of that, we have an implicit rule that prescribes explaining improbable occurrences in terms of causes and not in terms of effects (see §18).

Reichenbach treats the principle of common cause as a statistical problem. Assume that events A and B have been observed to occur frequently. Thus, with reference with a certain time scale, we can talk about the probabilities $P(A)$, $P(B)$ and $P(A, B)$. Consider now the following macrostatistical arrangement:

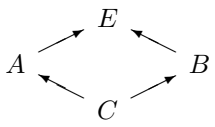

Figure 5: *Macrostatistical arrangement*

Fig.(5) is a double-fork arrangement in which A and B represent the two events the simultaneous occurrence of which is improbable, C is their common cause and E the common effect. The simultaneous occurrence of A and B is explained by C and not by E (Reichenbach argues in favour of a causal explanation rather than a final explanation in §18, see the example of the footprint traces in the sand). Because we are interested in explaining improbable occurrences in terms of causes and not in terms of effect, let us consider, in Fig. 6, only the lower part of the diagram in Fig. 5.

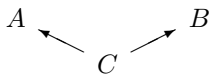

Figure 6: *Conjunctive fork*

Let us now examine the statistical relations holding for a single fork as in Fig.6. The simultaneous occurrence of A and B is more frequent than can be expected for chance coincidences:

$$P(A, B) > P(A) \cdot P(B) \tag{6}$$

Applying the multiplication theorem of probabilities to the left side we obtain:

$$P(A, B) = P(A) \cdot P(A|B) = P(B) \cdot P(B|A) \tag{7}$$

From (6), the two relations can be derived:

$$P(A|B) > P(B) \tag{8}$$

$$P(B|A) > P(A) \tag{9}$$

Each of these relations, vice versa, can be used to derive (6). Therefore, (8) or (9) is equivalent to (6).

Let us now see how the common cause is statistically characterized. Reichenbach assumes that the fork ABC satisfies the following relations:

$$P(C|A,B) = P(C|A) \cdot P(C|B) \tag{10}$$

$$P(\overline{C}|A,B) = P(\overline{C}|A) \cdot P(\overline{C}|B) \tag{11}$$

$$P(C|A) > P(\overline{C}|A) \tag{12}$$

$$P(C|B) > P(\overline{C}|B) \tag{13}$$

It can be shown that (6) is also derivable from these relations, therefore relations (10)-(13) define the conjunctive fork. The conjunctive fork, as Reichenbach says, is the statistical model for the principle of common cause. The rest of the paragraph is devoted to the statistical characterization of the conjunctive fork in order to define the direction of time in macrostatistics. Thus we now have the direction of time, that is the direction of time is the same as the direction of the causal relation. How about time ordering of events? That is, what is the time ordering of causes and effects?

Let us read §22. Macrostatistics, says Reichenbach, can be even used to define time order. The statistical relations mentioned above can replace the use of classical mechanics for the construction of a causal net which possesses a lineal order. We assume that events are observed as separate spatiotemporal units. To construct a causal net for events just in terms of statistical relations, events have to be classified into classes. To construct a causal net, two requirements are in order: (i) classes are codefined (a class A is codefined if it is possible to classify an event x as belonging to A coincidently with the occurrence of x); (ii) classes have a non 0 probability.

Now suppose we have three events x, y, z. This triplets of events occurs repeatedly and we wish to establish the causal ordering within each triplet. First, suppose we find out that x, y, z belong respectively to the classes A, B, C. To construct a causal net we need the relation of causally between. If we previously found out that that B is causally between A and C, then we say that y is causally between x and z. Assume we have a causal chain as in figure (7) If we know that x belongs to A, then we can predict with probability $P(B|A)$ that y will belong to B. Likewise, we can predict that C will occur from B. And here comes the awaited screening-off relation. Once we know that B occurred, A is no longer relevant for predicting C. As Reichenbach says, "the contribution of A

Figure 7: *A causal chain composed of three events*

to C has been absorbed in B, so to speak; and B may be said to *screen off* A from C". (1956, P. 189, my notation, emphasis in the original).

The relation causally between is defined through the following statistical relations:

$$1 > P(B|C) < P(A|C) > P(C) > 0 \tag{14}$$

$$1 > P(B|A) > P(C|A) > P(A) > 0 \tag{15}$$

$$P(A, B|C) > P(B|C) \tag{16}$$

And we symbolize this relation with $btw(A, B, C)$. It is worth noting that the relation causally between is symmetrical - *i.e.* given a three events arrangement, the two arrangements ABC and CBA are equivalment, but unique - *i.e.* given a three events arrangement, only one event will be causally between the other two (see Reichenbach 1956, §22 for details and proof). This means that because it is symmetrical causally between can't alone give us the time ordering yet, but because it is unique, we can use it for constructing the causal net.

For instance, suppose the following relations have been verified:

$$btw(A, B, C), btw(A, B, D), btw(D, B, C), \tag{17}$$

then we can conclude that the corresponding arragements is given as in figure (9) and not as in figure (8).

Figure 8: *A four events arrangement*

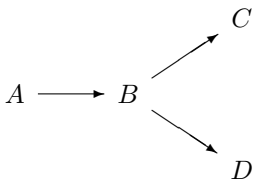

Figure 9: *A four event arrangement probabilistically equivalent*

However, Reichenbach is aware of the following difficulty: in the conjunctive fork the common cause as well as the common effect C can screen off A from B. In other words, the two arrangements in Fig. 10 and Fig. 11 are indistinguishable.

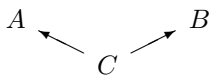

Figure 10: *Conjunctive fork: C is a common cause*

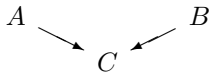

Figure 11: *Conjunctive fork: C is a common effect*

A couple of thoughts. As far as we consider relations among variables in a structural model comparable to macrostatistical processes, Reichenbach's reasoning about the conjunctive fork and screening off can be fruitfully applied in the social sciences. It is worth noting that, in the ultimate analysis, Reichenbach's screening-off relation has not a causal scope but rather temporal scope. Differently put, it is not a relation describing the "true" causal relations or what happens in the case of a third variable confounding the real causal effect, but it is just a temporal statistical relation to predict a further variable from temporally precendent ones. That is to say, from the fact that C screens off A from B in a conjunctive fork, it does not directly follow that we have to control for C.

4.2 Screening-off and control

The philosophical literature on causality and causal modelling makes extensive use of the term "screening-off", but no real conceptualization of it is put forward. "Screening-off" is generally used to show practical difficulties that causal modelling faces or the inadequacy of simple probabilistic characterization of the causal relations.

For instance, Suppes (1970) uses the screening-off relation for defining the genuine cause (a prima facie cause that is not spurious). In the second chapter Suppes presents a formal probabilistic theory of causal relations among events. Events are subsets of a fixed probability space, they are instantaneous and their times of occurrence are included in the formal characterization of the probability space (Suppes 1970, p. 12). It is worth noting that Suppes didn't mean the formal theory presented in chapter 2 to be the end of the story. As he says in the opening of chapter 2, he is rather formalizing the wide use of probabilistic causal concepts in ordinary talk. A sencond step will be to examine some systematic applications of the theory to various branches of science.

It seems to us that it is on this ground that Suppes' theory has to be evaluated. But let us first see how in his formal theory the "screening-off" relation is used.

A *prima facie* cause C is an event that, by definition, precedes the effect E in time and such that $P(E|C) > P(E)$ (assuming that $P(C)$ is non 0). Formally, the event $C_{t'}$ is a *prima facie cause* of E_t if, and only if:

1. $t^{'} > t$
2. $P(C_{t'} > 0)$
3. $P(E_t|C_{t'} > P(E_t)$

Of course, Suppes is aware of the fact that an earlier event, say F, may be found which accounts as well for the conditional probability of the effect. In this case C is a *spurious cause*. Let $C_{t'}$ be a prima facie cause; then, $C_{t'}$ is a *spurious cause* if and only if there is a $t^{''} < t^{'}$ and an event $F_{t''}$ such that:

$$P(C_{t'}, F_{t''}) > 0 \tag{18}$$

$$P(E_t|C_{t'}, F_{t''}) = P(E_t|F_{t''}) \tag{19}$$

$$P(E_t|C_{t'}, F_{t''}) \geq P(E_t|C_{t'}) \tag{20}$$

Suppes doesn't use graphical representation but supposedly the case of the spurious cause would appear as follows. $C_{t'}$ being a prima facie cause of E_t is represented by the graph in Fig (12). The

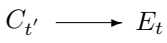

Figure 12: *Prima facie cause*

presence of an earlier event $F_{t''}$ that screens-off $C_{t'}$ from E_t can be represented as in Fig.(13), where the dashed arrow [it should be dashed!!] between $C_{t'}$ and E_t indicates that the relation is now spurious.

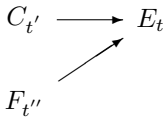

Figure 13: *Spurious cause*

Suppes is not particularly sure about condition (19) above, which he might want to replace with the inequality $P(E_t|C_{t'}, F_{t''}) \leq P(E_t|C_{t'})$ (Suppes 1970, p. 23). Equation (19) is formally equivalent

to Reichenbach's screening-off relation. However, whilst for Reichenbach, the screening-off relation has primarily a temporal meaning, *i.e.* it is used for the prediction of a future event, in Suppes' definition it has a truly causal significance. Together, conditions (19) and (20) suggest that the event genuinely responsible for the occurrence of the effect E is F and not C.

Cartwright (1979) points out that statistical analyses of causation such as Suppes' have failed because they can't deal properly with cases of Simpson's Paradox, *i.e.* in cases in which associations between two variables which hold in a given population can be reversed in the subpopulation by finding a third variable which is correlated with both. What Cartwright is actually criticizing is that the probability raising requirement is just too simplistic. "A cause - says Cartwright - must increase the probability of its effects-but only in situations where such correlations are absent" (Cartwright 1979, p. 423). Consequently, she advances the following definition:

> C *causes* E if and only if C increases the probability of E in every situation which is otherwise homogenous with respect to E.

This can be the case if all other factors in the population are held fixed, *i.e.* if we control for every possible *confounding factor*.

Irzik (1986) probably gives the most accurate discussion of screening-off. In this paper he claims an equivalence between the philosophical notions of statistical relevance and screening-off and the notions of correlation coefficient and zero-partial correlation coefficient used in causal modelling. Statistical relevance (*i.e.* $P(E|C) \neq P(E)$) is equivalent to a non-zero correlation in the case of dichotomous variables; the vanishing of (first-order) partial correlation coefficient ($\rho_{X,Y|Z} = 0$) conveys the same idea of screening-off, *i.e.* the correlation between X and Y disappears when controlled for Z. Irzik argues against Ellett and Ericson (1983), who defends the following reductionist rule in causal modelling.

> **Rule:** If the correlation r_{XZ} between X and Z is high positive (or negative) and the partial correlation coefficient $r_{Xz|Y}$ between X and Z with Y "held constant" is zero, either (a) Y is an intervening variable - the causal effect of X on Z (or vice versa) operates through Y; or (b) Y is a common cause of X and Z - the correlation between X and Z is spurious.

In other words, $r_{XZ|Y}$ is sufficient to infer:

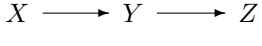

Figure 14: *Simple recursive system (a)*

or

Figure 15: *Simple recursive system (b)*

or

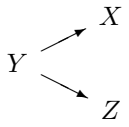

Figure 16: *Simple recursive system (c)*

However, as Irzik points out, Sewall Wright, as early as 1934 has shown that the following structure is note ruled out:

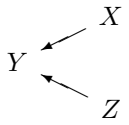

Figure 17: *Simple recursive system (d)*

Differently put, a zero-partial correlation does not allow us to distinguish Y as an intervening variable or as a common cause or as a common effect. Consequently, in line with Wright, Duncan and Blau, Irzik argues in favour of the non-reductionist character of causal modelling. In general, screening-off arguments are meant to defend a non reductionist position from statistics to causality. This is also the position held in Russo *et al.* (2006), where they point out that structural stability alone does not guarantee the causal interpretation of a structural conditional model. Well-founded field (or background) knowledge is indeed an essential element.

5 Heterogeneity, Latent exogenous variables, Frailty, Confounding variables and loss of exogeneity

We now examine some reasons possibly explaining why causal analysis may not be an easy task in current statistical modelling. We begin by considering a three-variates case and next extend the analysis to a p-dimensional vector. We first introduce a completely recursive decomposition and later consider more complex structures, in particular those arising from incomplete observability of

the relevant variables.

Thus, consider that, for data in the form $X = (Y, Z, U)$, the components of X have been so ordered that in the complete marginal-conditional decomposition :

$$p_X(x \mid \omega) = p_{Y|Z,U}(y \mid z, u, \theta_{Y|Z,U})\, p_{Z|U}(z \mid u, \theta_{Z|U})\, p_U(u \mid \theta_U) \tag{21}$$

each of the three components of the rhs may be considered as structural models with mutually independent parameters, *i.e.* in a sampling theory framework:

$$\omega = (\theta_{Y|Z,U}, \theta_{Z|U}, \theta_U) \in \Theta_{Y|Z,U} \times \Theta_{Z|U} \times \Theta_U \tag{22}$$

In such a case, (21) defines a *completely recursive system* and may be represented by Figure 18. This diagram suggests that U causes Z and (U, Z) cause Y. Also, equations (21) and (22) say that U is exogenous for $\theta_{Z|U}$ and that (U, Z) are jointly exogenous for $\theta_{Y|Z,U}$

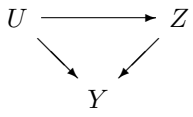

Figure 18: *3-component completely recursive system*

Now suppose that $Y \perp\!\!\!\perp U \mid Z$. In such a case, the structure (21) becomes:

$$p_X(x \mid \omega) = p_{Y|Z}(y \mid z, \theta_{Y|Z})\, p_{Z|U}(z \mid u, \theta_{Z|U})\, p_U(u \mid \theta_U) \tag{23}$$

Here, Z (alone) is exogenous for $\theta_{Y|Z} = \theta_{Y|Z,U}$. Figure 18 becomes Figure 19 and now U causes Z and Z causes Y; thus Z is "'endogenous"' in $p_{Z|U}$ and exogenous in $p_{Y|Z}$. Furthermore, if the process generating U were modified, or put under control, its effect of Y would be intermediated through Z; thus, for the effect on Y, Z may be regarded as a "sufficient summary" of U and Z.

Figure 19: *3-component completely recursive system with conditional independence*

Now suppose that U is not observable. It may be tempting to collapse the diagram in Figure 18 into that of Figure 20. Formally, Figure 20 may be obtained by integrating the latent variable U out of (21):

$$p_{Y|Z}(y \mid z, \theta_{Y|Z}) = \frac{\int p_{Y|Z,U}(y \mid z, u, \theta_{Y|Z,U})\, p_{Z|U}(z \mid u, \theta_{Z|U})\, p_U(u \mid \theta_U)\, du}{\int\int p_{Y|Z,U}(y \mid z, u, \theta_{Y|Z,U})\, p_{Z|U}(z \mid u, \theta_{Z|U})\, p_U(u \mid \theta_U)\, du\, dy} \tag{24}$$

$$p_Z(z \mid \theta_Z) = \int p_{Z|U}(z \mid u, \theta_{Z|U})\, p_U(u \mid \theta_U)\, du \tag{25}$$

Figure 20: *2-component system*

Therefore:

$$\theta_{Y|Z} \;=\; f_1(\theta_{Y|Z,U}, \theta_{Z|U}, \theta_U) \qquad \theta_Z \;=\; f_2(\theta_{Z|U}, \theta_U) \tag{26}$$

Many social scientists would call this situation a case of "confounding variable" (some scientists would even call it "the"' case), namely that U is a confounding variable or a common cause (of Y and Z). Statisticians would call it "a case of mixture models" or an issue on "missing data" or (unobservable) heterogeneity, or frailty model or ... and philosophers would probably call it "an interesting case"... All in all, this situation calls for several remarks:

- In general, Z is not exogenous anymore because (26) shows that the parameter $\theta_{Y|Z}$ and θ_Z are, in general, not independent; indeed some components of $\theta_{Z|U}$ and of θ_U may be common to $\theta_{Y|Z}$ and θ_Z (see however next remark);
- the non-observability of U typically implies a loss of identification: the functions f_1 and f_2 are *not* one-to-one; thus Z might still be exogenous because potentially common parameters in $\theta_{Y|Z}$ and θ_Z might not be identified;
- One might also look for further conditions providing the exogeneity of Z.

A frequently used simplifying assumption is the sampling independence between Z and U:

$$Z \perp\!\!\!\perp U \mid \omega \tag{27}$$

This assumption implies that $\theta_{Z|U}$ is now written as θ_Z and Figure 18 becomes Figure 21 suggesting that U and Z jointly cause Y (without U causing Z).

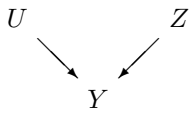

Figure 21: *3-component completely recursive system with marginal independence*

When U is not observable, Figure 20 is again obtained under the following integration of U:

$$p_{Y|Z}(y \mid z, \theta_{Y|Z}) \;=\; \int p_{Y|Z,U}(y \mid z, u, \theta_{Y|Z,U})\; p_U(u \mid \theta_U)\; du \tag{28}$$

Therefore:

$$\theta_{Y|Z} \;=\; f_3(\theta_{Y|Z,U},\, \theta_U) \tag{29}$$

independently of θ_Z and the exogeneity between Z and $\theta_{Y|Z}$ may be recovered; in particular, U is not anymore a common cause (of Y and Z) but, from (29), it is seen that the *meaning* of $\theta_{Y|Z}$ is to be a combination of the causal action of U along with that of Z, represented by $\theta_{Y|Z,U}$, and of the distribution of U, represented by θ_U.

Let us now see why the graphs represented in Figures 18 to 21 may be inappropriate to understand causal attribution. Firstly, causality is adequately represented in Figures 18, 19 and 21 once they represent structural models but in such cases Figure 20 is inappropriate for picturing causality because the underlying model represents solely a statistical model in terms of the manifest variables only. Secondly, Figure 20 is not appropriate to represent exogeneity of Z which is false in general but possible under the supplementary assumption (27). Last but not least, Figure 20 is not sufficient to interpret the statistical model, *i.e.* to understand the meaning of $\theta_{Y|Z}$, the parameter of $p_{Y|Z}(y \mid z, \theta_{Y|Z})$, and eventually to understand the effect of Z on Y; indeed these interpretations should be based on an explicitation of the functions f_1, f_2 and f_3.

An example may be useful to better grasp some difficulties. Suppose, for simplifying the argument, that the joint distribution of X in (21) is multivariate normal; thus the regression functions are linear and the conditional variances are homoscedastic, *i.e.* do not depend on the value of the conditioning variables. Let us compare the following two regression functions:

$$\begin{aligned}
\mathbb{E}\left[Y \mid Z, U, \theta_{Y|Z,U}\right] &= \alpha_0 + Z\alpha_1 + U\alpha_2 && (30)\\
\alpha_1 &= [cov(Y, Z \mid U)][V(Z \mid U)]^{-1} \\
&= [cov(Y,Z) - cov(Y,U)[V(U)]^{-1}cov(U,Z)] \\
&\qquad \times [V(Z) - cov(Z,U)[V(U)]^{-1}cov(U,Z)]^{-1} && (31)\\
\mathbb{E}\left[Y \mid Z, \theta_{Y|Z}\right] &= \beta_0 + Z\beta_1 \qquad \beta_1 = [cov(Y,Z)][V(Z)]^{-1} && (32)
\end{aligned}$$

Therefore, if the effect on Y of the cause Z is measured by the regression coefficient, the correct measure would be α_1 rather than β_1, once the conditional model generating $(Y \mid Z, U)$ is structural. Note that, in this particular case, $\alpha_1 = \beta_1$ when $Z \perp\!\!\!\perp U$ but this is a particular feature of the normal distribution for which $Z \perp\!\!\!\perp U$ implies that $cov(Y, Z \mid U) = cov(Y, Z)$ and $cov(Y, U \mid Z) = cov(Y, U)$ which is in general not true. Moreover, $\alpha_1 = \beta_1$ is also true when $\alpha_2 = 0$, *i.e.* when $Y \perp\!\!\!\perp U \mid Z$ which is contextually different from $Z \perp\!\!\!\perp U$.

This example makes several issues explicit:

(i) measuring the effect of a cause should be operated relatively to a completely specified structural model; failing to properly recognize this issue may lead to fallacious conclusions because in general: $\alpha_1 \neq \beta_1$

(ii) apparently ancillary specification, such as a normality assumption, may be more restrictive than first thought; for instance, under a normality assumption, the hypotheses $U \perp\!\!\!\perp Z$ and $Y \perp\!\!\!\perp U \mid Z$ implies each that $\alpha_1 = \beta_1$ although they are contextually different, once the normality assumption is not retained. This is so because, in the normal case, independence is equivalent to uncorrelatedness and the regression functions are linear.

In the previous analysis, Y and Z may be indifferently random variables (*i.e.* univariate) or random vectors. Let us now consider a decomposition of X into p components: $X = (X_1, X_2, \cdots X_p)$. Suppose that the components of X have been so ordered that in the complete marginal-conditional decomposition :

$$\begin{aligned} p_X(x \mid \omega) \quad = \quad & p_{X_p|X_1,X_2,\cdots X_{p-1}}(x_p \mid x_1, x_2, \cdots x_{p-1}, \theta_{p|1,\cdots p-1}) \\ & \cdot p_{X_{p-1}|X_1,X_2,\cdots X_{p-2}}(x_{p-1} \mid x_1, x_2, \cdots x_{p-2}, \theta_{p-1|1,\cdots p-2}) \cdots p_{X_1}(x_1 \mid \theta_1) \quad (33) \end{aligned}$$

each components of the rhs may be considered as structural models with mutually independent parameters, *i.e.* (in a sampling theory framework):

$$\omega = (\theta_{p|1,\cdots p-1}, \theta_{p-1|1,\cdots p-2} \cdots, \theta_1) \in \Theta_{p|1,\cdots p-1}) \times \Theta_{p-1|1,\cdots p-2} \cdots \times \Theta_1 \quad (34)$$

The representation of (33) through the chain (35):

$$X_1 \to X_2 \to \cdots \to X_i \cdots \to X_p \quad (35)$$

might be viewed as a simplified representation of a completely recursive system where, for instance, the first two arrows would actually stand for Figure 22. The first three arrows would be represented as in Figure 23 and so on.

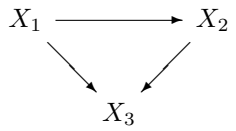

Figure 22: *First 3 components of a completely recursive system*

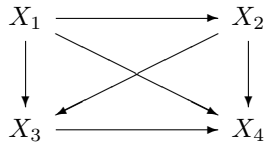

Figure 23: *First 4 components of a completely recursive system*

This is however a dangerous road. Once the value of p increases such graph representations become quickly unmanageable *unless* some simplifying assumptions, in the form of conditional independences, operate simplifications of the same nature as Figure 21 simplifies Figure 18. This is indeed a basic issue in structural modelling: the use of field knowledge aims not only at ordering the components of X, so as to obtain (33), but also at bringing in more structure than the complete system (33). The simple 3-variable case has however suggested how complex may be to assess the role of further assumptions.

More specifically, the statistical modelling of a complex systems raises several issues:

1. Given a p-dimensional vector of variables to be modelled, is the field knowledge sufficient for ordering the variables in such a way that one may obtain a completely recursive system as in (33), *i.e.* in such a way that each component X_j is univariate ? As a matter of fact, if often happens, in particular in econometrics, that it is not possible to disentangle recursively the process generating a vector of variables, in other words that some components X_j are subvectors of X rather than univariate random variables. For instance, Mouchart and Vandresse (2005) handles a case where the data are made of vectors the components of which are price and attribute of a set a contracts concluded through a bargaining process: the data and the contextual information do not allow to know whether the prices have been bargained after or before the attributes have been agreed upon. This is a case of *simultaneity* where the model describes a process generating a vector of (so-called "endogenous") variables conditionally on a vector of exogenous variables, in such a way that the equations of the model do not correspond to a marginal-conditional decomposition. The econometric literature, particularly between the sixties and the eighties, is extremally rich in developing this class of models, called "simultaneous equations models".

2. Endowing each distribution of (33) with a structural interpretation amounts to say that each of these distributions represents a contextually relevant data generating process. The principle of parsimony recommends to focus the attention on the processes of actual interest and is made operational by selecting a subvector $(X_{r+s}, X_{r+s-1}, \cdots, X_r)$ of X such that the joint distribution of $(X_{r+s}, X_{r+s-1}, \cdots, X_r \mid X_1, \cdots X_{r-1})$ gathers all data generating processes of actual interest. In such a case the subvector $(X_1, \cdots X_{r-1})$ becomes globally exogenous for the system of interest.

3. Quite a difficult issue in structural modelling is bound to the fact that many theories, in social sciences, involve latent, or nonobservable, variables introduced in order to help structuring a theoretical construct; think, for instance, to the concept of "intelligence" in psychology, of "social easiness" in sociology or of "permanent income" in economy. In such a case, a structural

model is built including latent and manifest, or "observable", from which a statistical model is obtained by integrating out all the latent variables. A typical benefit of such an approach is to obtain a statistical model with more structure, *i.e.* more restrictions, than a "saturated" statistical model constructed independently of a structural approach. A well-known case is provided by the LISREL type model, or covariance structure model. However that structural approach has also a cost, sometimes difficult to handle. Indeed, the analysis performed around the simplest case of one unobservable variable along with two observable variables, given through equation (24) and (25) suggests that the analysis of exogeneity *at the level of the statistical model bearing on the manifest variables only* becomes soon untractable, jeopardizing most exogeneity properties and making difficult the interpretation of the identifiable parameters.

6 Control

6.1 Ex ante at the Data Level

If possible confounding is suspected, it should ideally be taken into account at the design stage of the study (see *e.g.* Rothman and Greenland 1998, Lee 2005). The best way to avoid confounding bias in a prospective study is ex ante randomisation, *i.e.* a random allocation of subjects between the 'treatment' and the 'control' groups. The first group receives the treatment and the second a placebo. Moreover, in view of avoiding investigator bias, the procedure should be kept secret from the investigator. Actually, double-blind experiments are often conducted in clinical epidemiology, both the investigator and the subject ignoring whether the latter is put into the treatment group or into the control group. Randomisation ensures to a high extent that both groups differ only by the fact that one receives the treatment (the 'cause') and the other not. The causal relation between treatment and outcome (*e.g.* recovery) is not affected by a confounder - observed or unobserved - due to the fact that the two groups are similar in all respects except treatment/no treatment. In particular there is no selection bias, though there might be a placebo effect in the no treatment group. This does not mean however that the target group is homogeneous as concerns treatment effect; the group may contain sub-groups which vary widely in response to the treatment. In non-random allocations, there is always a risk that the characteristic on which the allocation is done is associated with the outcome so that the treatment effect cannot be distinguished from the selection effect, though matching can to some extent reduce the problem, Lee (2005). On the other hand, the variables the influences of which are randomised are included in the disturbance term; this increases the disturbance variance and makes it more difficult to discover an experimental effect, Bollen (1989, p. 74).

Though quite common in clinical studies, randomisation can however be unethical or impossible

to conduct. For moral reasons, one may not randomly allocate subjects, for example, between a smoking group and a non-smoking group, or between receiving a treatment A and no treatment at all if an alternative treatment is already available. Presently, in clinical trials, the alternative treatment must be given to the control group instead of a placebo. To give another example, allocating subjects to different income groups in order to examine the causal relation between income and fertility would not be unethical but impractical. For these reasons, randomisation is very often impossible in the social sciences. It is however practised in educational studies, among others, in order to compare *e.g.* the performance of students following a physics course based on lectures with a physics course based on project development. In addition, to avoid course-teacher interaction, teachers are usually rotated between both types of courses.

In retrospective studies, ex ante randomisation is not possible, as the subjects either have or have not been subjected to the causal factor and/or to the outcome (Holland and Rubin 1988). A case-control study on the relation between smoking and lung cancer would compare the past smoking history of persons alive with lung cancer, *i.e.* those having experienced the outcome, with a control group of ex post randomly selected persons without lung cancer (Khlat, 1994 and the other articles in the special issue of this journal on case-control studies). Furthermore, several controls are usually matched to each case on some possible major confounders such as age or gender. Matching does not ensure however that cases and controls are alike on other variables. All possible sources of confounding bias are not avoided, contrary to ex ante randomisation in prospective studies. In addition, the comparison between cases and controls can be done in terms of odds ratios but not of relative risks, as the population exposed to risk is unknown. Finally, only persons alive are included in the study and they may differ from those who have died. On the other hand, retrospective case-control studies are much less expensive and time-consuming to carry out than prospective randomised trials. Case-control studies are too rarely conducted in the social sciences.

6.2 Ex Post Stratification

In non-experimental (observational) studies, prospective or retrospective, the two major ex post approaches for controlling for confounders are stratification for categorical variables and statistical adjustment for numerical variables. This distinction is not clear-cut however, as statistical adjustment can be applied to categorical variables using e.g. logit regression, and a numerical variable can always be categorised - age can be transformed into age groups for example. These approaches can take into account observed confounders but are hardly able to control for latent ones. Only the stratification approach will be developed here; for other methods, see *e.g.* H. Wunsch *et al.* (2006).

In an observational study, stratification implies conditioning on the confounding variable(s). This approach will be considered in the case of the well-known Simpson's paradox; the example is taken from Pearl (2000, pp. 174-175). In a population suffering from a disease, one sub-population follows

a treatment and the other does not. The treatment increases the recovery rate when both genders are combined. On the other hand, when gender is taken into account (controlled for), the drug decreases the recovery rate both for males and for females, thus the paradox. The following table (Table 1) distributes the population by treatment, recovery, and gender.

Both genders	**Recovery**	**No recovery**	**Total**	**Recovery rate**
Treatment	20	20	40	.50
No treatment	16	24	40	.40
Total	36	44	80	
Males				
Treatment	18	12	30	.30
No treatment	7	3	10	.70
Female				
Treatment	2	8	10	.20
No treatment	9	21	30	.30
Total	11	29	40	
Source: Pearl, 2000				

Table 1: Treatment, recovery, and gender

It is easy to show why the treatment seems to have an effect in the general population while this conclusion does not hold in each gender category. The combined recovery rates in the Treatment and No treatment groups can be written as follows:

$$(30x0.60 + 10x0.20)/40 = 0.50 \text{ for the Treatment group}$$

$$(10x0.70 + 30x0.30)/40 = 0.40 \text{ for the No treatment group}$$

In the Treatment group, there is a higher proportion of males than females, and males have a higher recovery rate than females. The opposite is true in the No treatment group. Males thus opt for treatment more than females and they have a higher recovery rate because, for example, their compliance might be higher. Gender is therefore a confounding factor, as the two sub-populations differ by gender structure (*i.e.* , gender has an impact on opting for treatment or not) and gender influences the recovery rate.

In order to obtain an unbiased global indicator, gender combined, standardisation is often recommended: the rates by gender are applied to a same arbitrary standard population structure. This procedure actually blocks in this example the causal link from gender to treatment use. Blocking the link from the confounder to either the cause or the outcome is sufficient for controlling for the confounder. For example, taking the Treatment population structure as standard, one would obtain the following global rates:

$$(30x0.60 + 10x0.20)/40 = 0.50 \text{ for the Treatment group}$$

$$(30x0.70 + 10x0.30)/40 = 0.60 \text{ for the No treatment group}$$

This time, the combined standardized rate does lead to the same results as the analysis by gender. The recovery rate (both genders) is lower in the treatment group than in the recovery group; furthermore, no interaction is present here as the absolute difference is the same for males and for females, *i.e.* 10 per cent. In this situation, the same global result would be obtained whatever the standard population. The treatment should therefore be discontinued.

However, standardisation breaks down when there are interaction effects (see G. Wunsch (2006), for a general discussion). In the case of strong interaction, no linear combination of rates can represent the true results because more than one indicator per group is required in this situation. If this were the case in the present example, no averaging of the results by gender could lead to a satisfactory global indicator. For example, consider the situation where the recovery rates would be 0.60 (males) and 0.40 (females) in the treatment group, and respectively 0.70 and 0.30 in the No treatment group. According to the choice of the arbitrary standard population, one could say that the global recovery rate in the Treatment group is lower, equal, or higher than in the No treatment group, a conclusion which is not very informative! No sole measure can tell us that the recovery rate for males is lower in the treatment group compared to the No treatment group while the converse is true for females (i.e. strong interaction between treatment and gender).

Suppose now that the hospital, where the patients are treated, is associated both with treatment and recovery. In addition to taking a sample of patients per hospital, one can also draw a sample of hospitals and take into account both the treatment effect at the individual level and the hospital effect at the contextual level, using a multilevel model. Daniel Courgeau has shown that this type of model elegantly resolves Simpson's paradox; it can furthermore include interaction terms between the individual and contextual variables, see Courgeau (2002), Courgeau (2003). The approach lies however outside the scope of this paper.

7 When to control or not to control

When should we control for a background variable and when should we not control for this variable, when examining the impact of a treatment/exposure on an outcome in an observational (*i.e.* . non-experimental) study? In other words, is this variable a confounder or not? If the background variable is a common cause of both treatment/exposure and outcome, it should be controlled for. Let us go back to the previous figures 2 and 4. In Fig. 2, SES has to be controlled for in this model if one wants to estimate the impact of smoking on cancer of the respiratory system in the absence of confounding, taking into account the definition of a confounder given in the previous section. SES

is a common cause of smoking (exposure) and of cancer (outcome) via A, as SES is a parent of both tabacism and asbestos exposure. In this case, conditional on SES, the variables tabagism and asbestos exposure should become independent and the association between them should disappear; if not, a latent confounder(s) is most probably present. To give another example, in the model of figure (4), if one is interested in the relation between T and C, A should be controlled for as A is a common cause of both the exposure T and the outcome C. Suppose now that SES is itself latent, *i.e.* unobserved. We could also block the impact of SES by controlling for A on the path from SES to C (the outcome), as A is observed, or by conditioning on an observed intermediate variable between SES and T (the treatment/exposure). In the absence of such observed intervening variables, we could also control for any observed variable K, *e.g.* income, deemed to be highly correlated with the latent variable SES on the basis of our background knowledge, using K as a surrogate for SES , Hernán *et al.* (2002). One's conceptual framework, based on background knowledge, theory, and research hypotheses, should take into account for this purpose all known observed and latent variables relevant for the problem at hand (Gérard, 2006).

Actually, even if SES is observed in the model of figure (2, it would still be advisable to control for A instead of for the common cause SES if we want to measure the impact of smoking on cancer. SES might not be the only common cause of T and A, as pointed out in the previous paragraph. For example, smoking and occupation are gender-dependent; gender would also be a common cause. To give another example, an unknown gene G may make some smokers and asbestos inhalers more susceptible to cancer; the effects of smoking and asbestos exposure would then be associated through the common genotype. The common cause might also be too remote from the causal relations studied. For example, one could argue that one's SES is dependent upon one's parents' SES and control for the latter. However, the more remote, the less influence the common cause probably has on the variables downstream in the model. Therefore, we would usually recommend controlling for more proximate intervening variables on the path from a common cause to the outcome than on the common cause itself if the latter is much further upstream in the causal graph, especially in the social sciences where multiple latent common causes are probably the norm. Note that in this case one falls back on the classic definition of a confounder as a variable associated with the treatment/exposure (through the presence of a common 'ancestor') and having an impact on the outcome!

Controlling for a variable that is not a confounder can be harmful. In the model of figure (3), A should not be controlled for when studying the total impact (direct and indirect) of T on C, as A is not a confounder but an intermediate variable in the indirect path going from T to C through A, in addition to the direct path T to C. Controlling for A would only be justified if one wished to evaluate solely the direct effect of T on C.

Furthermore, controlling for a common effect of a treatment and an outcome creates a spurious

association between the latter two, Hernán *et al.* (2002), in addition to a possible treatment effect. This is due to the fact that two causes become correlated when one controls for their common effect (or collider). As an example, Hernán considers two independent variables - diet and non-diet-related cancer - which become associated once we control for weight loss, a common effect of both diet and cancer. One should therefore not condition on a common effect of two (or more) variables.

Another situation where it would be inappropriate to control for a covariate is the case of a *conjunction of causes* of the type $XZ \rightarrow Y$, *i.e.* X and Z jointly cause Y. To give an example, in the case of fertility transition, Ansley Coale (1973) considers that several necessary conditions (conscious choice, advantage, effective techniques) have to be simultaneously satisfied for fertility to fall in high fertility countries. If at least one of these conditions is not satisfied, fertility will not decrease: for a decline in fertility to occur, the conjunction of the three conditions is required. A causality of conjunctions of causes has been proposed in philosophy by John Stuart Mill (1889) and Richard Taylor (1966), and by John Mackie (1974) with his INUS causality.

Take the very simple situation where the three variables X, Z, Y, are dichotomous (presence/absence). One obtains in this case the following results

X	**Z**	**Y**
1	1	1
0	1	0
1	0	0
0	0	0

Controlling for Z (conditioning on Z) means examining the relation between X and Y for $Z = 1$ or $Z = 0$. If $Z = 1$, one would conclude that X is a necessary and sufficient cause of Y because "if X , then Y" and if "non X, then non Y". On the contrary, if $Z = 0$, one would conclude that X is not a cause of Y because $Y = 0$ for both $X = 0$ and $X = 1$. The general conclusion would be the existence of an interaction effect between X and Z as the influence of X on Y varies according to the values of Z. Actually, this would be a misspecified model; in reality, the interaction between both causes is so strong that the additive effects of X and of Z on Y disappear in favour of the sole conjunction XZ influencing Y.

We said above that it is probably better to control for an intervening variable between the common cause and the effect, rather than for the common cause itself especially if the latter is remote from the treatment/exposure, due to the possible presence of other - latent - causes or the temporal reduction of influence for remote common causes. Here is however a case, borrowed from Pearl (2000) and taken up by Greenland and Brumback (2002), where this strategy would be wrong. In this model, all the variables are observed. Consider five variables U, V, Z, X, and Y, causally

related as in the directed graph of figure (24), and suppose we are specifically interested in the relation X causes Y.

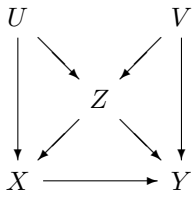

Figure 24: *Common causes and intervening variable*

In this model, contrary to intuition, it is not sufficient to control for the intervening variable Z. Even if U and V are independent, they are associated in some strata in Z as they both cause Z. Due to this association, U is associated with Y through V and V is associated with X through U. Either V or U must be controlled for to remove the confounding.

Finally, common causes also have an impact on the way we deal with competing risks. For example, in demography it is customary to estimate a net probability of dying at a given age in the absence of migration using a counterfactual approach. If migrants had not left the country, what would have been the probability of dying without the occurrence of migration? One usually assumes that both processes are independent, leading to the well-known Berkson formulas (see *e.g.* G. Wunsch 2002). Suppose now that a common cause such as marital status influences both mortality and migration, which is actually the case. The two variables are then associated, as Reichenbach's conjunctive forks have shown. In this case, migration could select persons with different mortality probabilities than non-migrants and the classic formulas for computing net probabilities would be biased. Once again, the counterfactual approach is hardly acceptable except as a very simplified model; it would be better to control for the common cause(s) if possible.

8 Conclusions

Structural models give a meaningful framework for defining causality. Those models, congruent with contextual knowledge and characterized by parameters that are stable over a large class of interventions, allow us to approach causality in terms of exogeneity in a structural conditional model. It is in *this* framework that we analyse the issue of confounding and of control.

The presence of confounding factors, *i.e.* of third factors that screen-off one variable from another, has been been discussed at length in modelling conjunctive forks by Hans Reichenbach. The philosophical literature has generally followed up this tradition and argued that, since confounders raise fundamental problems for discovering causal relations, confounders ought to be always controlled for. A closer look at Reichenbach's modelisation of conjunctive forks reveals, however, that

the concept of screening-off has predictive compass rather than a causal one. In fact, in ordering a sequence of events in time, say A, B and C, knowing A is no longer relevant for predicting C once we know B, and thus we say that B screens-off A from C. The same probabilistic structure applies to conjunctive forks, where the common cause screens-off one effect from the other. Thus, once the common cause is taken into account - or controlled for - the correlation between the two effects vanishes.

Confounders are common causes of both treatment/exposure and of response/outcome. Some classical definitions of confounding in epidemiology or in demography, based on associations only, are incomplete as they do not consider the causal paths among the variables which lead to association. Confounding is better taken care of by randomization at the design stage of the research. Randomization is however unethical or impossible to achieve in many social science problems. In observational studies, conditioning on the observed confounders by stratification is recommended but the results should not be standardized if there is strong interaction between the confounder and treatment/exposure on the outcome. If the confounder is latent, controlling for observed intervening variables or an observed surrogate confounder may sometimes be possible. Even when the common cause is observed, it might be better in some cases but not in all to control for an intervening factor if other latent common causes are suspected. The research strategy should be based on a thorough knowledge of the field and on one's conceptual framework.

In this structural framework, the concepts of exogeneity and of causality have been explicitly defined; however, we have shown that the impact of latent variables complicate substantially the analysis and the operational interpretation of those concepts.

Acknowledgment The research underlying this paper is part of a research project conducted by the three authors, on Causality and Statistical Modelling in the Social Sciences. Parts of this paper have been prepared for the seminar on Causality, Exogeneity and Explanation, Causality Study Circle - Evidence Project, UCL, London, May 5, 2006. Financial support to M.Mouchart from the IAP research network nr P5/24 of the Belgian State (Federal Office for Scientific, Technical and Cultural Affairs) is gratefully acknowledged. F. Russo acknowledges the support of ...(FR has to mention FSR funding)...(to be completed). The authors would like to thank Daniel Courgeau, Myriam Khlat, Laurent Toulemon and Marie Vandresse for their comments, suggestions, questions and doubts on various sections of this paper.

References

Anderson S., Auquier A., Hauck W.W. , Oakes D., Vandaele W., and Weisberg H.I. (1980), *Statistical Methods for Comparative Studies*, Wiley, New York.

Barndorff-Nielsen O. (1978), *Information and Exponential Families in Statistical Theory*, New York: John Wiley.

Bartecchi C.E., Mackenzie T.D., Schreir R.W. (1995), The global tobacco epidemic, *Scientific American*, 272(5), 26-33.

Bollen K.A. (1989), *Structural Equations with Latent Variables*, New York: John Wiley & Sons.

Cartwright N. (1979), Causal laws and effective strategies, *Nous*, **13**, 419-417.

Coale A.J. (1973). The demographic transition reconsidered, *Proceedings of the International Population Conference*, Volume 1, IUSSP, Liége, 53-72.

Courgeau D. (2002). Réflexions sur Régression et analyse géométrique des données, Henry Rouanet *et al.* , unpublished (personal communication).

Courgeau D. (2003), From the macro-micro opposition to multilevel analysis in demography, chapter 2 in D. Courgeau (Ed.), *Methodology and Epistemology of Multilevel Analysis*, METHODOS Series N 2, Kluwer, Dordrecht, pp. 43-91.

Cox D.R. and Wermuth N. (2004), Causality: a statistical view, *International Statistical Review*, 72(3), 285-305.

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

Ellett F. and Ericson D. (1983), The logic of causal methods in social science, *Synthese*, **57**, 67-82.

Elwood J.M. (1988), *Causal Relationships in Medicine*, Oxford University Press, Oxford.

Engle R.F., Hendry D.F. and Richard J.-F. (1983), Exogeneity, *Econometrica*, **51**(2), 277-304.

Fisher R.A. (1957), Alleged dangers of cigarette smoking, *British Medical Journal*, II, 297-298.

Fisher R.A. (1958), Lung cancer and cigarettes, *Nature*, 182, July 12, p. 108.

Florens J.-P. and Mouchart M. (1985), Conditioning in Dynamic model, *Journal of Times Series Analysis*, **53** (1), 15-35.

Florens J.-P., Mouchart M. and Rolin J.-M. (1990), *Elements of Bayesian Statistics*, New York: Marcel Dekker.

Freedman D. (1999), From association to causation: some remarks on the history of statistics, *Statistical Science*, 14(3), 243-258.

Gaumé C. and Wunsch G. (2003), Health and Death in the Baltic States, in I. E. Kotowska and J. Jozwiak (eds.), *Population of Central and Eastern Europe. Challenges and Opportunities*, Statistical Publishing Establishment, Warsaw, 301-325.

Gérard H. (2006), Theory building in demography, chapter 129 in G. Caselli, J. Vallin, and G. Wunsch, *Demography. Analysis and Synthesis*, Volume 4, Academic Press, San Diego, 647-660.

Greenland S. and Brumback B. (2002), An overview of relations among causal modelling methods, *International Journal of Epidemiology*, 31, 1030-1037.

Hernán M.A., Hernández-Diaz N., Werler M.M., and Mitchell A.A. (2002), Causal knowledge as a prerequisite for confounding evaluation: an application to birth defects epidemiology, *American Journal of Epidemiology*, 155(2), 176-184.

Holland P.W. and Rubin D.B. (1988), Causal inference in retrospective studies, *Evaluation Review*, 12(3), 203-231.

Hood Wm. C. and Tjalling C. Koopmans (editors) (1953), *Studies in Econometric Methods*, Cowles Foundation Monograph 14, New Haven: Yale University Press.

Irzik G. (1986), Causal modelling and the statistical analysis of causation, *PSA 1986*, 12-23.

Jenicek M. and Cléroux R. (1982), *Epidémiologie*, Maloine, Paris.

Kesteloot H., Sans S., and Kromhout D. (2006), Dynamics of cardiovascular and all-cause mortality in Western and Eastern Europe between 1970 and 2000, *European Heart Journal*, 27(1), 107-113.

Khlat M. (1994), Use of case-control methods for indirect estimation in demography, *Epidemiologic Reviews*, 16(1), 124-133.

Lee M-J (2005), *Micro-Econometrics for Policy, Program and Treatment Effects*, Oxford University Press, Oxford.

Leridon H. and Toulemon L. (1997), *Démographie. Approche statistique et dynamique des populations*, Economica, Paris.

Mackie J.L. (1974), *The Cement of the Universe: A Study of Causation*, Clarendon Press, Oxford.

Mill J.S. (1889), *A system of logic*, Longmans, Green and Co., London.

Mouchart M. and A. Oulhaj (2000), On Identification in Conditional Models, Discussion paper DP0015, Institut de statistique, UCL, Louvain-la-Neuve (B).

Mouchart M. and Vandresse M. (2005), Bargaining Power and Market Segmentation in Freight Transport, to appear in *Journal of Applied Econometrics*.

Oulhaj A. and M. Mouchart (2003), The Role of the Exogenous Randomness in the Identification of Conditional Models, to appear in *Metron*.

Pearl J. (2000), *Causality*, Cambridge University Press, Cambridge.

Reichenbach H. (1956), *The direction of Time*, University of California Press. Reprinted (2000), Dover Publications, Mineole, N.Y.

Robins J.M. (2001), Data, design, and background knowledge in etiologic inference, *Epidemiology*, 11(3), 313-320.

Rothman K.J. and Greenland S. (1998), *Modern Epidemiology*, 2nd edition, Lippincott-Raven, Philadelphia.

Rouanet H. (1985), Barouf á Bombach, *Bulletin de Mthodologie Sociologique*, 6, 3-27.

Russo F., M. Mouchart, M. Ghins and G. Wunsch (2006), Statistical Modeling and Causality in Social Sciences, Working paper submitted for publication.

Schlesselman J.J. (1982), *Case-Control Studies - Design, Conduct, Analysis*, Oxford University Press, New York.

Spirtes P., C. Glymour and R. Scheines (2000), *Causation, Prediction and Search*, second edition, Cambridge (Mass.): The MIT Press.

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

Taylor R. (1966), *Action and Purpose*, Prentice-Hall., Englewood Cliffs, NJ.

Vallin J., Caselli G., and Surault P. (2006), Behavior, lifestyles, and sociocultural factors of mortality, chapter 51 in G. Caselli, J. Vallin, and G. Wunsch, *Demography. Analysis and Synthesis*, Volume 2, Academic Press, San Diego, 143-170.

Wunsch G. (2002), The life table: a demographic overview, in G. Wunsch, M. Mouchart, and J. Duchêne (Eds.), *The Life Table. Modelling Survival and Death*, Kluwer, Dordrecht, 13-31.

Wunsch G. (2006), Confounding variables, standarization, and the problem of summary indices, chapter 15 in G. Caselli, J. Vallin, and G. Wunsch, *Demography. Analysis and Synthesis*, Volume 1, Academic Press, San Diego, 197-208.

Wunsch H., Linde-Zwirble W.T., and Angus D.C. (2006), Methods to adjust for bias and confounding in critical care health services research involving observational data, *Journal of Critical Care*, 21(1), 1-7.