Potential Outcomes, Counterfactuals, and Structural Modelling

Causal Approaches in the Social Sciences

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

[a] *Institut Supérieur de Philosophie, UCLouvain, Belgium*

& Philosophy, University of Kent, UK.

[b] *Institut de Démographie, UCLouvain, Belgium.*

[c] *Institut de Statistique, UCLouvain, Belgium.*

May 26, 2008

Preliminary Version

Not to be quoted

Comments welcome

Abstract

This paper examines the potential outcome model developed by Rubin and its counterfactual underpinnings as developed by Lewis. Though a major contribution of Rubin's potential outcome model has been to stress the importance of the design stage, we recall the main methodological and epistemological flaws of his approach. We argue that the study of causes and effects does not necessarily require counterfactuals, once a structural modelling framework, as the one developed here, is adopted. Our approach emphasises and spells out the role of background knowledge, marginal-conditional decomposition, and of stability for providing a causal explanation of a given phenomenon.

Keywords: Causality, Counterfactuals, Potential outcomes, Structural Modelling.

Corresponding Author: Federica Russo, Institut Supérieur de Philosophie, UCLouvain, Belgium, Place Mercier 14, B-1348 Louvain-la-Neuve , Belgium. e-mail: federica.russo@uclouvain.be

Contents

1 Introduction

One goal of social science, though not the only one, is to find out causes of phenomena, to measure effects of causes, and to provide causal explanations. The search for causes and effects has a long tradition both in the scientific and philosophical debate. The idea that to establish causal relations we have to check what happens or what would happen were the putative cause be absent rather than present is not new. Some philosophers, notably David Lewis (1973a), argued that this was already implied by the definition of cause Hume (1748) gave in the *Enquiry Concerning Human Understanding*. After Hume, this very same idea was at the very basis of the Millian experimental methods (Mill 1843). Whilst Lewis developed this idea in the context of causal language by providing a formal analysis for subjunctive conditionals, a formal counterfactual analysis has been proposed in the statistical literature only around the 1970s in the pioneering work of Donald Rubin (1974).

Rubin's model has been very influential but also severely criticised from many quarters. We argue here that criticisms against the potential outcome model are indeed sound, but that they go only half way through. We then propose a general framework based on structural modelling as an alternative to the potential outcome/counterfactual approach. Our answers are articulated throughout the paper as follows. In this paper, we tackle the two following questions: (i) Are all the criticisms addressed to the potential outcome model sound? (ii) Are counterfactual questions to be dismissed altogether? In a nutshell, we answer yes to the first question and no to the second one.

Section 2 presents Rubin's potential outcome model. Rubin's model measures the causal effects of treatments, e.g. the effect of aspirin on headache, assuming that for each individual in the study we can potentially observe both the outcome of the treatment, e.g. taking aspirin, and the outcome of its alternative, e.g. not taking aspirin. The causal effect is then the difference between the two potential outcomes for the individual considered. An average causal effect for the population is then computed from the individual effects. Although it is applied to nonexperimental situations, Rubin's approach strongly reflects the treatment group versus control group design in randomised experiments. This section also presents traditional criticisms and basically agrees with them, especially with the problem of not observing at the same time for the same individual the outcomes of the two possible exposures. In addition, the potential outcome model measures the effects of causes but is ill-suited for uncovering the causes of effects. This section also points to the fact that according to Rubin, non manipulable factors cannot be considered as *causal* factors in the model. We argue that this is a major flaw in his account because gender or ethnicity, for instance, may be significant determinants of income or HIV/AIDS respectively. It is worth noting that because one of the two potential outcomes is unobservable, the potential outcome model leads to formulate contrary-to-fact questions, also known as counterfactuals.

Section 3 explores the original counterfactual account developed by the philosopher and logician David Lewis. The reason why we go into the details of this account is that present-day counter-

factualists in statistics and in social science often claim that their ideas originate in those of Lewis. In the paper "Causation", Lewis (1973a) presented an account where causal relations are analysed in terms of subjunctive conditionals, also known as *counterfactuals*. A causal statement such as 'A caused B' is then interpreted as 'B would not have occurred if A had not occurred'. The peculiarity of those claims is that they are conditional statements the antecedent of which is known to be false. Due to the paradoxes of the material implication, classical propositional logic cannot regiment counterfactuals exactly because they would all be equally true, given that the antecedent is false. For this reason a different sematics, also known as possible world semantics, has been developed. We then argue that the "statistical" and the "philosophical" accounts, although both dealing with counterfactual statements, have different scopes, aims and applications. On this ground we distinguish between "statistical" and "philosophical" counterfactuals, and we draw further distinctions between single-case, individual, and generic causal statements—these are meant to clarify the meaning of various causal statements. Lewis' counterfactuals aim at detecting single-case causal relations as used in everyday language, e.g., 'Had *Mr Jones* taken an aspirin half an hour ago, his headache would have gone now'. But they do not aim at uncovering generic causal relations—e.g., 'Aspirin relieves headache'—nor at measuring individual (yet generic) causal effects of treatments—e.g., 'The causal effect of aspirin on an individual randomly sampled from the population is such and such'.

A main assumption behind the potential outcome model is that we assume what the cause of a given effect is – for instance, we assume that aspirin relieves headache and based on this assumption we construct and evaluate counterfactual statements. The model is therefore ill-suited for uncovering the causes of effects. Furthermore, the potential outcome model is not apt for studying the various paths, direct and indirect, leading from the cause(s) or treatment(s) to the effect(s) or response(s), and more generally for examining the network of relationships among the variables. A major problem, in addition, is the need for controlling for assignment bias, i.e. for the fact that in nonexperimental situations the assignment of units to the treatment and control groups is often the result of self-selection. As Rubin himself has stressed, well formulated causal models are needed in the social sciences because controlling for the relevant covariates may not be trivial without a properly developed causal model.

In section 4, inspired by the seminal works of Wright, Haavelmo, Blalock, Pearl and others, we develop a structural modelling approach to causation. In essence, a model is deemed structural if it uncovers a structure underlying the data generating process. This approach systematically uses three ingredients. First, the model must be congruent with background knowledge: modelling the data generating process must be operated in the light of the current information on the relevant field. Second, the structure is expressed by an ordering of the relevant variables and decomposition of the joint distribution into an ordered sequence of conditional distributions. Third, the model must show stability in a wide sense: both the structure of the model and the parameters have to

be stable or invariant with respect to changes of contexts. It is crucial to note that this concept of structural modelling is wider than the framework of structural equations models, also known as covariance structure models or LISREL type models, widely used in psychology or in sociology, and of simultaneous equations models, widely used in econometrics. Finally, we discuss both the structural modelling and the potential outcome model and argue that counterfactuals can make sense and be legitimately applied only if based on sound structural modelling.

2 Rubin 's potential outcome model

2.1 Rubin's definition of a causal effect: counterfactuals and potential outcomes

Consider the classic case of a person who is treated at time t. To be simple, the outcome or response to the treatment is observed at time $t+k \quad (k>0)$. How does one conclude that the treatment is effective or not? In other words, how do we measure the possible causal effect of the treatment? Donald Rubin's answer to estimating the causal effect of treatments in randomized and nonrandomized studies is based on a counterfactual proposition such as: "If an hour ago I had taken two aspirins instead of just a glass of water, my headache would now be gone" (Rubin, 1974). Following Rubin's notation, if E represents taking two aspirins and C drinking just a glass of water, the potential outcomes Y relating to these two treatments may be written as two random variables, namely $Y(E)$ and $Y(C)$. The causal effect of the E versus C treatment on Y for a particular subject j observed at times t and $t+k$ is then defined as $Y_j(E)-Y_j(C)$, *i.e.* the differential headache response to taking the aspirins or not taking them.

If we consider N subjects instead of only one, one has a causal effect $Y_j(E)-Y_j(C)$ per subject j. The average causal effect for this group of N persons can then be written

$$\frac{1}{N}\sum_{1\leq j\leq N}[Y_j(E)-Y_j(C)] \tag{1}$$

Rubin's solution is often called the *potential outcome (or response) model*, the two potential outcomes being in this simple case $Y_j(E)$ and $Y_j(C)$ for each j. Note that the causal effect may differ from one individual to the other, thus a "typical" causal effect (Rubin's term) is obtained as above by taking the average (or any other summary measure) of the individual causal effects. As pointed out by Brand and Xie (2007 p.394), "the potential outcome approach to causal inference extends the conceptual apparatus of randomized experiments to the analysis of nonexperimental data, with the goal of explicitly estimating causal effects of particular "treatments" of interest".

In the actual world, one never observes at the same time for the same individual both $Y(E)$ *and* $Y(C)$. The subject either takes (or is assigned to) E or to C. Thus one can never observe for a same individual j the causal effect $Y_j(E)-Y_j(C)$. In general, people are assigned either to E or to C but

not to both at the same time. Suppose however that one wishes to measure the use-effectiveness of an IUD (intra-uterine device) versus an oral contraceptive (pill) on pregnancy outcome. Usually, $N/2$ women would be assigned to the IUD and $N/2$ to the pill, and their eventual fertility compared. Women could nevertheless be assigned to both treatments at the same time, *i.e.* to the IUD and to the pill. The differential responses would then be $Y(IUD\ and\ pill)$ versus $Y(IUD)$ or $Y(pill)$ in addition to $Y(IUD)$ versus $Y(pill)$.

Still following Rubin (1974), suppose there are only two subjects under study, denoted by 1 and 2. The typical causal effect (as defined above in the counterfactual situation) would then be

$$1/2[Y_1(E) - Y_1(C) + Y_2(E) - Y_2(C)]. \tag{2}$$

In the actual world, one would observe in a single study either

$$Y_1(E) - Y_2(C) \tag{3}$$

or

$$Y_2(E) - Y_1(C) \tag{4}$$

depending on whether subject 1 or subject 2 is assigned to E, and vice versa 2 or 1 to C.

If treatments are randomly assigned to subjects, we are equally likely to observe the difference (3) or (4). The expected difference in the outcome Y is then the average of equations (3) and (4):

$$1/2[Y_1(E) - Y_2(C)] + 1/2[Y_2(E) - Y_1(C)] \tag{5}$$

It is easily seen that under randomization, equation (5) is equal to equation (2). In other words, (5) is an unbiased estimate of 2).

Suppose now that subjects 1 and 2 respond similarly to the treatments E and C In that case

$$Y_1(E) - Y_2(C) = Y_2(E) - Y_1(C) \tag{6}$$

and furthermore

$$Y_1(E) - Y_2(C) = Y_1(E) - Y_1(C) \tag{7}$$

or

$$Y_2(E) - Y_1(C) = Y_2(E) - Y_2(C) \tag{8}$$

In the situation of perfectly matched subjects with respect to the effects of the treatments, the observed causal effect is equal to the counterfactual causal effect. Results under randomization or perfect matching can be extended from two subjects to N subjects. Randomization and matching are therefore two approaches to measuring the causal effect in experimental and nonexperimental studies, though randomization cannot often be used in the social sciences and perfect matching is hardly possible in practice. In many actual situations in nonexperimental research, the assignment

of units to the case and control groups is based on self-selection. Thus the assignment procedure is often not "ignorable", in the sense that the likelihood of treatment on the one hand and the outcome on the other hand are not independent. For example, if the sickest opt for the new treatment and the healthier for the older one, the outcome (e.g. recovery) in the treatment group will be due both to the new drug and to the characteristics of the patients at onset. In this case, one must control as best as possible for the assignment factors which have an impact on the outcome. In the above example, one would try to control for the state of health of both groups at the beginning of the trial.

It should be noticed that Rubin requires that all subjects have to be potentially exposable to either E or C, *i.e.* to the various k treatments $(E_1, E_2, E_3, ..., E_k)$ - including possibly no treatment - being compared. In this approach, "causes are only those things that could, in principle, be treatments in experiments" (Holland, 1986). Therefore, an attribute (such as gender or ethnicity) cannot be a cause because potential exposability does not apply to it. In other words, in this framework there is "no causation without manipulation". For example (Rubin, 1986), a study on gender differences in starting salaries cannot be addressed by randomized experiments and therefore gender cannot be a cause of differential salaries among subjects. Gender is an attribute and cannot be considered in the search of effects of causes. According to Rubin, there is no clear causal answer to this issue.

2.2 Epistemological flaws

A major contribution of Donald Rubin's potential outcome model has been to stress the importance of carefully planning the design stage in observational studies. In particular, the assignment mechanism by which some units are subjected to the putative cause ("treatment" group) and others not ("control" group) should be studied in depth prior to any data analysis of the outcomes, and thoroughly explicated if possible. "We should objectively approximate, or attempt to replicate, a randomized experiment when designing an observational study" (Rubin 2007). For this purpose, Rubin has developed propensity score methods destined to eliminate bias, by setting up subclasses such that within each subclass, the treatment and control units have similar distributions on the known covariates influencing assignment. This approach requires however that the assignment mechanism is otherwise unconfounded, i.e. it assumes that there are no latent confounders influencing the assignment of units between the treatment and control groups. This assumption is not required in experimental studies where the units are assigned randomly to the treatment and control groups.

Though Rubin's potential outcome model is a significant contribution to analysing the cause-effect relation in observational studies, it nevertheless suffers from some important epistemological flaws which are examined now.

Potential outcomes: a "Platonic heaven?" A major criticism that has been addressed to Rubin's potential outcome (or potential response) model is its counterfactual basis (Dawid, 2007). Paul W. Holland (1986) has even called it the *fundamental problem of causal inference* . The individual causal effect, as proposed by Rubin, requires taking the difference $Y_j(E) - Y_j(C)$, though one of the two potential outcomes will never be observed. "There is no world, actual or conceivable, in which both variables could be observed together. Their simultaneous existence must therefore be confined to some "Platonic heaven" of ideal forms, not fully accessible to real-world observation" (Dawid 2007, p. 510). Actually, taking an assignment variable X, let $X = 1$ denote the fact that subject j is assigned to treatment E while $X = 0$ denotes assigning the *same* subject j to the other treatment C. The latter can be a placebo for example. Then for *the same subject j at the same time t* , the probability $P(X = 1|X = 0) = 0$ and vice versa $P(X = 0|X = 1) = 0$. In other words, if John Smith is assigned to the treatment E, his probability of being assigned to the placebo C at the same time is nil (and vice versa).

It follows that one cannot assume two outcomes, one corresponding to E and the other to C for the same John Smith. One of the potential outcomes is not only unobservable: its occurrence is impossible, because the same subject cannot be assigned both to E and to C at the same time, according to the potential exposure assumption. "E and C are exclusive of each other in the sense, that a trial cannot simultaneously be an E trial and a C trial " (Rubin, 1974).

Attributes A major failing of the potential outcome model is that it cannot take attributes into account (Ni Bhrolchain and Dyson, 2007). *Pace* Rubin, gender *is* a cause of initial salary discrimination in many countries, ethnicity *is* a cause of differential HIV prevalence in Sub-Saharan Africa, etc. These attributes are not only associated with their respective effects - they are part of the causal mechanism itself. For example, ethnic groups in Africa have different reproductive norms, values, and sexual behaviors (such as multi- or single-partnership), and these characteristics are major determinants of exposure to HIV. Any explanatory framework in the social sciences that cannot take attributes into account is therefore necessarily flawed. The statement "no causation without manipulation" is much too strong in this case, and weaker assumptions should be considered in order to take attributes too into account.

Morgan and Winship (2007 p.280) have countered this argument by evoking the construction of thought experiments. For example, "the counterfactual model could be used to motivate an attempt to estimate the average gain an employed black male working full time, full year would expect to capture *if all prospective employers believed him to be white*" (italics ours). However, there exists an 'infinity' of possible thought experiments for each case and no way of testing the validity of their claims with actual data. In the previous example, one could estimate the difference in income between blacks and whites controlling if possible for all income factors other than race

(such as level of education, health status, etc.). No hypothetical counterfactual thought experiment is actually required here. The real problem is both knowing and observing the factors which have to be controlled for.

Some authors such as Paul Holland and James Woodward (see Woodward 2003, chapter 2) contend that the issue in the gender/salary example is actually not to manipulate gender, but in this case to modify the beliefs concerning gender, or the attitudes and practices of the employer as to hiring females, *i.e.* variables that can be manipulated contrary to gender. Though correct, this proposal can nevertheless not be extended to all the cases of attributes as causes. Consider the example of sex (male, female) as a major risk factor of breast cancer. No manipulation of the patient's or the physician's beliefs and attitudes towards breast cancer will change the fact that breast cancer is about 100 times less common among men than among women. The biological differences between males and females explain this relation, though the cause cannot be manipulated in practice. The problem of integrating attributes into the manipulation theory of causation therefore still remains unresolved. In our view, the *concept* of causality should not be dependent upon conditions of manipulability.

Causes of effects The potential outcome model largely derives from experimental models where units are randomly assigned to disjoint sets of treatments. It focuses on the 'effects of cause' problem and can hardly tackle the 'causes of effect' issue which is central to much of the social sciences (Ni Bhrolchain and Dyson, 2007). Though favouring a counterfactual approach to causality himself, Heckman (2005, p.2) has nevertheless pointed out that "Science is all about constructing models of the causes of effects", and insists on the need to understand the causes producing the effects, or in other words the determinants of the outcomes.

Though randomization has indeed proved very useful as a method enabling to distinguish causal effects from non-causal ones, randomization is by no means an essential element of the *concept* of causality. As Heckman (2008) has stressed: "The claim that causality can only be determined by randomization reifies randomization as the 'gold standard' of causal inference". In the social sciences, randomized experiments are often difficult to conduct but nevertheless causal patterns have been discovered in all disciplines in the absence of randomized experiments, by a careful control of the relevant covariates and by using criteria supportive of causal inference (Ni Brolchain and Dyson op. cit.). As Rubin (1974) himself has stressed, more well formulated causal models are needed in the social sciences because controlling for relevant covariates may not be trivial without a properly developed causal model. In their recent book, Morgan and Winship (2007) opt for a counterfactual approach but they actually deal mainly with causal modeling in the spirit of Pearl (2000), which does not necessarily require counterfactual assumptions, *pace* Pearl.

The individual or the population? Statistics is a methodology of learning by observing within the framework of a heterogeneous population of reference. As Ronald Fisher wrote a half-century ago already, "The conception of statistics as the study of variation is the natural outcome of viewing the subject as the study of populations" (Fisher, 1958). Rubin's focus on the *individual* causal effect and the average of individual effects runs counter to the fact that statistics yields population effects and not individual ones. What we can say *e.g.* of our own individual probability of survival is inferred from the population life table. Similarly, causal effects can only be obtained at the population level, as counterfactual questions at the individual level are unanswerable. Maybe it would have been better, an hour ago, if I had taken two aspirins instead of just a glass of water, but I will never know. On the other hand, population studies have shown that the probability of relieving one's headache is higher after taking aspirins than after drinking water. Next time I have a headache, I'll try taking aspirins instead of just drinking water...

Counterfactuals or causal modeling? In his influential book on causality, Pearl (2000) distinguishes between two languages for causality that have been proposed: path analysis/structural equation modeling on the one hand and the Neyman-Rubin potential outcome model on the other hand. We have seen that for various reasons the potential outcome model, as applied to specific individuals, does not seem very convincing. It is indeed impossible to answer counterfactual questions at the level of the individual. This does not mean however that counterfactual questions should be dropped from the causal language. On the contrary, it is common practice at the population level to raise the question 'if the putative cause had not occurred, would the effect have occurred'? This query is one of several criteria used as aids for causal attribution in epidemiology for example (Beaglehole, Bonita, Kjellström 1994, chapter 5) and it is also at the basis of the treatment group (cause present) / control group (cause absent) comparison. The major problem here is assignment bias and confounding of the cause - effect relation by other variables. To solve these problems, as Rubin has stated, well-developed causal models are necessary, taking into account the network of variables related to the cause and to the effect.

Following Pearl's distinction of the two causal languages recalled above, and in agreement with Fienberg's argument for representing every quantity under consideration using random variables and displaying them in directed acyclic graphs (Fienberg 2006), a structural modeling approach to causation will be proposed in section 4 as an alternative to the counterfactual approach. Structural modeling should not be confused with structural equation modeling, which is but one methodology among others. Actually, as we shall see, structural modeling need not be restricted to quantitative methods only; qualitative methodology is often better suited both for exploratory research and in-depth knowledge. Before tackling the structural modeling approach, the following section will first examine Lewis' counterfactual conditionals and possible worlds semantics. The defenders of the

counterfactual approach (see e.g. Brand and Xie op. cit.) have indeed often related their approach to the philosophy of David Lewis and more particularly to his book *Counterfactuals* (1973b). We show however that the two theories are different and should not be assimilated.

3 Lewis' counterfactuals

3.1 Counterfactuals and possible worlds semantics

Famously, David Hume (1748, sec.VII) defined a cause as

> [...] an object followed by another, and where all the objects similar to the first are followed by objects similar to the second. *Or, in other words, if the first object had not been, the second had never existed* (italics ours).

David Lewis, in his pioneering works (Lewis 1973a and 1973b), thought that the second part of the well-known definition given by Hume was not just a restatement of the first claim, but a clear encouragement to think of causality in *counterfactual* terms. But what is a counterfactual exactly? For an excellent overview of the history, problems, and prospects of counterfactuals in the philosophical literature see the *Introduction* in Collins, Hall and Paul (2004). In the following, we shall content ourselves with offering to the reader some basic notions about Lewis' counterfactuals, enough to grasp the philosophical origins of the counterfactual models in quantitative causal analysis and the objections moved to them.

A counterfactual is a subjunctive conditional statement the antecedent of which, *i.e.* , in general, the first part of the statement, states a contrary-to-fact situation. Consider again the aspirin example: "Had Mr Jones taken an aspirin half an hour ago, his headache would have gone now". This conditional statement presupposes that Mr Jones did not take the aspirin and still has headache. Conversely, had he instead taken the aspirin he wouldn't have headache anymore. This type of subjunctive conditional statements are also called *counterfactuals*. It can easily be shown that classical propositional logic does not fit the case of counterfactuals. In fact, if we were to analyse subjunctive conditionals as simple material implications of the form $A \rightarrow B$—which reads "if A *is* true, then B *is* true—given that the antecedent is false, all counterfactuals would be equally true. This is the case because in classical propositional logic any material implication ($A \rightarrow B$) is true whenever the antecedent is false or the consequent true ($\neg A \vee B$). This means that the formulae $A \rightarrow B$ and $\neg A \vee B$ are logically equivalent. Therefore, a logical analysis of counterfactuals has to be run on a different ground, namely we have to find meaningful truth conditions for the counterfactual conditional. In the Seventies, due especially to the works of Robert Stalnaker (1968) and David Lewis (1973a), a *possible-world semantics* for counterfactuals has been developed.

Simply put, possible-world semantics is based on modal logic, *i.e.* , the logic that deals with the notions of possibility and necessity. In Lewis' account, possible-world semantics rests on the assumption of the existence of a plurality of worlds, among which there is also our actual world. This position is also known as modal realism. Modal realism is, needless to say, a metaphysical problematic position. However, letting aside all the problems it raises, we can think of possible worlds simply as "alternative situations" or "states of affairs". For the purpose of the present discussion, such simplification will do. The idea is that if a proposition is *possibly* true, this means that there must be at least one situation, or world, in which the proposition is true. Conversely, if a proposition is *necessarily* true, then the proposition has to be true in all conceivable situations or worlds, including of course the actual world. Thus, to evaluate a counterfactual conditional, we need to know in which worlds the antecendent and the consequent are true. Worlds are compared with each other on the basis of their similarity or closeness, and ranged according to their similarity. To order worlds, we use a relation of *comparative over-all similarity* which is taken as primitive:

> A world w_1 is closer to our actual world w_a than another world w_2 if w_1 resembles to w_a more than w_2 does.

The truth of the counterfactual is then ascertained by an "inspection" of what happens in other possible worlds. Let us now introduce a minimum of formalism. Given any two propositions A and B, the counterfactual $A\Box \rightarrow B$ reads: "if A *were* true, then B *would* also be true". The counterfactual operator $\Box \rightarrow$ is defined by the following rule of truth:

> The counterfactual $A\Box \rightarrow B$ is true (at a world w_i) if, and only if:
>
> 1. there are no possible A-worlds (where A-world means "the world in which A is true"), or
> 2. some A-world where B holds is closer to w_i than is any world where B does not hold.

The second case is the interesting one, as in the former the counterfactual is just vacuously true. Notice, however, that, in case A is true, the A-world is just our actual world and $A\Box \rightarrow B$ is true if, and only if, B is. This, in a nutshell, is the logics regimenting counterfactual statements. Let us illustrate with the following example: "If I were to drop my pen, it would fall on the floor". Intuitively, this counterfactual is true, given that the law of gravity holds: in fact, a world in which I drop my pen and it doesn't fall on the floor would be much further away from the actual world where the law of gravity actually holds.

What about causation? Lewis (1973a) develops an account where causal relations are analysed in terms of counterfactuals. "A caused B" is interpreted as "B would not have occurred if it were not for A". Causation comes in because

> We think of a cause as something that makes a difference, and the difference it makes must be a difference from what would have happened without it. Had it been absent, its effects—some of them, at least, and usually all—would have been absent as well. (Lewis 1986, p.160-161)

In other words, everything else being equal, a world in which Mr Jones takes the aspirin and the aspirin doesn't relieve his headache is further away from another possible world in which Mr Jones takes the aspirin as well and indeed the aspirin does relieve his headache. Why? Well, because this is what aspirin is supposed to do. But, needless to say, the question is how do we know whether, in general, aspirin relieves headache or not? Lewis does not tackle this question, while Rubin and other counterfactualists in social science do.

It is worth noting that Lewis states the scope of his analysis clearly (Lewis 1986, p.161-162). First, his discussion covers causation among *events*, in the everyday sense of the word: causation is a relation between events. For instance, the event "taking the aspirin" causes the event "recovering from headache". However, counterfactuals are *propositions*. But such a linguistic analysis can be easily applied to events because, although presumably events are not propositions, according to Lewis, they can at least be paired with them (Lewis 1986, p.166). Obviously events are not the only thing that can cause or that can be caused, but Lewis' original account, nor his latest account (Lewis 2004), goes beyond that. Second, his analysis is meant to apply to singular cases and not to generalisations. The distinction between generic and single case will be thoroughly dealt with later in subsection 3.3. For now, it will be enough to make clear that singular causal relations concern specific events that actually occurred, whilst generic causal relations try to generalise from a number of instantiated cases or to extrapolate a causal relation that is valid for the population of interest. An example will clarify. One thing is to ask whether had *Mr Jones* taken the aspirin, his headache would have gone now. Another thing is to ask whether aspirin is an effective treatment for headaches and therefore, given any individual randomly sampled from the population, aspirin would relieve or would have relieved his/her headache. In the first case we are dealing with a single-case causal relation, whereas the in the second case we are dealing with a generic causal relation.

Let us revert to Lewis' analysis again. Causality, recall, comes in because by asking whether the counterfactual $A\Box \rightarrow B$ is true, we wonder whether B would be a consequence of the occurrence of A, *i.e.* , whether the occurrence of A is the *cause* of the occurrence of B. So, the counterfactual states, if true, that is if the cause had not occurred, then the effect would not have occurred either. This condition is also called *counterfactual dependence*. However, counterfactual dependence cannot, alone, be a sufficient condition for causation for at least two further qualifications are needed (Lewis 2004): (i) the kind of relata, and (ii) the kind of counterfactual conditionals.

In fact, as for the relata, we need causes and effects to be distinct events, that is non-identical events that do not overlap and do not imply each other. As for the choice of the counterfactual

conditional to evaluate, it is worth noting that counterfactuals work under the so-called *ceteris paribus* conditions, that is, everything else being equal, the effect would not have occurred had the cause not occurred. The choice of the counterfactual conditional rests on what we decide to hold fix. To borrow Lewis' persuading example, imagine Caesar in command in Korea in the Fifties, then what we can hold fix is either Caesar's military knowledge or the weaponry used in the Korean war. The choice of one or the other will lead to different counterfactual conditionals. So, although the notion of comparative similarity of possible worlds is taken as primitive, it constitutes altogether a problematic aspect for the account, as there is no objective and unique ordering of possible worlds.

3.2 Traditional criticisms and difficulties

Everyday causal language makes extensive use of counterfactuals, thus we can rightly say that counterfactuals do grasp part of the meaning of what it is for an event to cause another event. Nonetheless, counterfactuals face a variety of problems. Again, Collins, Hall and Paul (2004) provide an excellent overview and reconstruction of the problems in Lewis' account, but see also Menzies (2001) for a discussion of cases of failure of transitivity, preemption, and chancy causation. As Collins, Hall and Paul (2004) say, it is worth distinguishing between genuine counterexamples to Lewis' account and challenges meant to better specify its foundations.

For instance, of the former types are examples showing failure of transitivity of the causal relation or of preemption. Although we usually think causation to be a transitive relation, that is if A is a cause of B and B is a cause of C, then A is a cause of C, there might be cases in which the corresponding *counterfactual dependence* relations do not hold. Another problematic case for counterfactual is that of "preemption". A typical example discussed in the literature is the case of two assassins willing to kill the same person by different methods and the action of the first "preempts", that is it makes causally inefficient, the action of the second. "Prevention" also troubles the counterfactual account: how are counterfactualists going to analyse cases where an event prevents another event to occur?

Consider now the challenges. The whole point about the counterfactual account to causation is that we have to investigate what would have happened, had the putative cause not occurred. Though intuitively simple, this can indeed be a tricky job, for events might be more fragile or more poorly defined than we think. Consider the aspirin example again: "Had Mr Jones taken the aspirin half an hour ago, his headache would have gone now". But what if Mr Jones went for a walk, or took paracetamol instead, or consulted a holy man? What if he had taken the aspirin later rather than sooner? How different would this event be with respect to the original antecedent? Lewis was aware of this kind of problem but was not particularly clear as how to solve it in his seminal 1973 account nor in his latest 2004. This kind of problem leads straight away to the urgency of an account of events: if causation is a relation between events, we need to know what they are in the first place,

and it is far from being trivial to produce a good theory of events.

This leads to a related problem, also called "context-sensitivity", that is the sensitivity of causal relations to contextual factors. This aspect is overlooked in Lewis' theory, although a fairly clear distinction between causes and conditions is available in the literature since the seminal work of Hart and Honoré in 1985.

Finally, Lewis' account works under the assumption of determinism. Although he treated the probabilistic case, his probabilistic account rests problematic altogether. In the probabilistic case, says Lewis, the occurrence of the antecedent event A makes the occurrence of the consequent event counterfactually more probable. The counterfactual then reads as: had the cause not been, the chance of occurring of the effect would have been much less than it actually was. Here probability is interpreted as temporally indexed single-case chances, which raises of course philosophical problems about the interpretation of probability.

The foregoing discussion just points to some difficulties of Lewis' account without pretending to be exhaustive nor to offer solutions. The moral to be drawn so far is that Lewis' goals were to analyse causal relations in terms of counterfactual conditionals, and that the causal relations he was interested in primarily were relations between occurred events, that is single-case causal relations and not generic ones.

3.3 Lewis' counterfactuals vs Rubin's counterfactuals: single-case vs generic

Many counterfactualists, both in the statistical and social science literature, trace the origins of the ideas behind the counterfactual approach in the work of Lewis. For instance, Pearl even claims a formal equivalence between Lewis' account and his account (Pearl 2000, ch.7). In the following, our goal will not be to run an exegetic investigation of Rubin, nor to question the formal equivalence claimed by Pearl. The goal will not be to dismiss counterfactual reasoning either. Counterfactual reasoning does play a major role both in everyday causal reasoning and in scientific reasoning. What we aim at showing next is that in spite of a strong analogy, Lewis' counterfactuals and Rubin's counterfactuals do not have the same scope, that it to say, they do not aim at establishing the same sort of causal claims. In a nutshell, whilst the former concern *single-case* causal relations, the latter concern *generic* causal relations.

Consider our usual example of aspirin and headache. On the one hand, the potential outcome model might want to establish whether aspirin is an effective treatment for headache, namely whether *aspirin relieves headache*. Of course, the fundamental quantity is the *individual* causal effect, that is the effect for the unit being treated minus the effect for the united not being treated. Surely this concerns single cases, that is the individual causal effect is measured using *individual* data. However, the whole point about the potential outcome model is *not* whether or not I or Mr Jones

would have recovered had I or he taken an aspirin, *but* rather whether aspirin is an effective treatment in the target population. Therefore, given any individual randomly sampled from the population, everything else being equal, her/his headache would go, were s/he to take an aspirin. On the other hand, Lewis, as explained above, asks what the truth conditions of a counterfactual statement are. Therefore he asks, given a particular situation—e.g., Mr Jones has been suffering from headache for the last four hours, and had he taken an aspirin he would feel good now—whether the counterfactual claim picks out the right cause.

True, the analogy is definitively there—Rubin's counterfactual exploits the same idea behind Lewis' counterfactuals: had the cause not been, the effect would not occur either, but this does not imply that these accounts be the same or that their scope be the same. Moreover, we are now pushed to draw the distinction between generic and single-case more clearly.

As we have seen, in philosophy, and particularly in Lewis' account, the counterfactual approach was motivated by the Humean definition of cause—if the cause had not been, the effect would have never existed either. According to Lewis, *singular* causal relations are established by means of an evaluation of counterfactual statements. If we want to know whether taking the aspirin actually relieved Mr Jones' headache, or whether it would have relieved his headache had he took it, we have to ascertain the truth of the corresponding counterfactual statement. The kind of causal relation we are here evaluating is single-case, namely a specific causal relation taking place at a certain time and place. We are not evaluating the causal effectiveness of aspirin in relieving headache in a target population, which is exactly the purpose of the potential outcome model.

As discussed earlier in section 2, Rubin uses the aspirin example as a single-case relation. However, this is a wrong interpretation of statistical counterfactuals. Granted, results of a counterfactual model can and are to be applied to single cases (*e.g.* diagnosis or causal attribution) but what we contend is that the scope of the counterfactual model developed by Rubin be single-case. Also, it is true that Rubin's potential outcome model and more generally counterfactual models use *individual* data, but this doesn't mean that they concern *individual* or single-case causal relations directly. The result of a counterfactual model would sound like this: more often than not, taking aspirins relieves headache, therefore, given any individual randomly sampled from the population, had s/he taken the aspirin, his/her headache would have gone. This is not the same as saying that 'had *Mr Jones* taken the aspirin, his headache would have gone now'. The former counterfactual, although based on individual-level data, is *generic*, whilst the latter is *single-case*, that is it concerns a particular causal relation taking place in a given time and place.

4 Structural modelling: a general framework

Due to the criticisms that can be addressed to the counterfactual model, in this section we develop the idea of structural models, *i.e.* models uncovering the structure underlying the observed phenomena and providing a causal explanation. We first recall the nature of a statistical model (Section 4.1) and next the concept of a structural statistical model (Section 4.2). We explain the meaning of structural modelling through the process of model building, with special attention to how using and incorporating background knowledge. In a nutshell, this involves developing a conceptual model out of background knowledge and then translating it into an operational model taking into account the indicators available, in the spirit of H.M. Blalock (1971) and, more recently, H. Gérard (2006). Finally we point out that the structural modelling approach may include non-statistical models.

We also emphasise that our approach differs from other uses of "structural" (*e.g.* , structural equation modelling as in Heckman (2005), or Pearl's causal graphs as in Pearl 2000 and Halpern and Pearl (2005)) in that (i) it is not characterised by a particular class of statistical models, (ii) it is rooted into the process of accumulating knowledge, and (iii) it is not based on counterfactuals.

4.1 Statistical Models

Let us first recall that, formally, a statistical model $\mathbf{M}$ is a set of probability distributions, explicitly:

$$\mathbf{M} = \{\mathbf{S}, \mathbf{P}^{\omega}\ \omega \in \mathbf{\Omega}\} \tag{9}$$

where S, called the sample space or observation space, is the set of all possible values (or, the range space) of a given observable (random) variable (or vector of variables) and for each $\omega \in \Omega$, P^{ω} is a probability distribution on the sample space, also called the sampling distribution. Thus, ω is a characteristic, also called parameter, of the corresponding distribution and Ω describes the set of the sampling distributions belonging to the model.

Roughly speaking, two ways of building a statistical model may be distinguished. A first one derives the model from observed associations and other descriptive properties of a given body of data. Such models are called descriptive models, associational models, or in some contexts exploratory data analysis, data mining *etc.* An alternative approach is to look for an underlying data generating mechanism, namely an underlying structure. Such models may be called structural models and are going to be the object of the following analysis. The basic idea here is that the data be analyzed, and explained, as if they were a realization of one of those distributions, characterised by Ω. Such a statistical model is accordingly based on a stochastic representation of the world. Its randomness delineates the frontier or the internal limitation of the statistical explanation, since the random component represents what is not explained by the model (Mouchart, Russo and Wunsch, 2008).

4.2 The meaning of "structural"

Suppose we want to evaluate, or predict, the effect of increasing family allowances on the fertility of the female. A descriptive strategy would look for empirical associations among changes in family allowances and changes in fertility, examining data from different periods and/or from different countries, possibly taking into account some covariates and considering the goodness of fit as a standard for the quality of the model. Now, if we want to understand the process generating fertility and the possible effect of increasing family allowances, we should try to uncover the underlying structure of the data generating process. With this objective, we should incorporate, in our strategy for model building, the most relevant available knowledge we have on the phenomenon. In particular, we should incorporate biological knowledge; for instance, the effect on fertility should not be expected before 9 months after taking the policy measure, nor should we expect an effect on women outside the fertility period of their life. We should also incorporate knowledge from psychology, sociology and, possibly, economics. All this information constitutes our background knowledge, also designated as knowledge of the field.

That this background knowledge contributes to the explanation of the relations being investigated means that, under an appropriate ordering of the involved random variables, there is the possibility of operating a recursive decomposition of the multivariate process (Cox and Wermuth, 2004). This is a decomposition of the multivariate distribution into the product of a sequence of conditional distributions, each term being conditional on an increasing sequence of the components of the random vector and such that each component is congruent with background knowledge.

More explicitly, let us now consider a decomposition of X into p components, namely $X = (X_1, X_2, \cdots X_p)$, and suppose that the components of X have been ordered in such a way that in the complete marginal conditional decomposition:

$$\begin{aligned} p_X(x \mid \omega) \quad &= \quad p_{X_p|X_1,X_2,\cdots X_{p-1}}(x_p \mid x_1, x_2, \cdots x_{p-1}, \theta_{p|1,\cdots p-1}) \\ &\quad \cdot p_{X_{p-1}|X_1,X_2,\cdots X_{p-2}}(x_{p-1} \mid x_1, x_2, \cdots x_{p-2}, \theta_{p-1|1,\cdots p-2}) \cdots p_{X_1}(x_1 \mid \theta_1) \quad (10) \end{aligned}$$

each component of the right hand side may be considered, in a first step, as a structural model with mutually independent parameters, *i.e.* in a sampling theory framework:

$$\omega = (\theta_{p|1,\cdots p-1}, \theta_{p-1|1,\cdots p-2} \cdots, \theta_1) \in \Theta_{p|1,\cdots p-1} \times \Theta_{p-1|1,\cdots p-2} \cdots \times \Theta_1 \quad (11)$$

Equations (10) and (11) characterize a *completely recursive system*. A recursive decomposition is not complete when, in equation (10), some components are random vectors rather than random variables. This typically happens when we cannot order some of the variables, due to a lack of knowledge on their causal or temporal priority. In such a case, there is, in the factorization (10), (at least) one factor giving as structural the distribution of a set (or, a vector) of variables, say X_j, conditional on the antecedent ones $(X_1, \cdots, X_{j-1})$. This situation is known under the heading of "simultaneity" in

the econometric literature (or "system with feedback", in the engineering literature). An example in epidemiology is discussed below.

In a causal perspective, a (completely) recursive system can generally be simplified. Indeed, after ordering the variables, some conditioning variables are deemed not to be causal on the basis of background knowledge and they may be dropped from some of the factors of the rhs of (10): the remaining variables are then postulated causal in their relative factors. This corresponds to deleting arrows (directed edges) from the associated directed acyclic graph, on the basis of conditional independence arguments. In other words, causality is viewed as exogeneity in a structural conditional model, considered in the framework of a (not always completely) recursive decomposition.

As an example, consider the relation between socioeconomic status (SES) and cancer of the respiratory system (C). Different SES categories present different cancer mortality/morbidity risks. Is SES a cause of C? What are the mechanisms involved? Background knowledge tells us that exposure to tabacism (T) and to asbestos (A) varies among socioeconomic groups. Furthermore, both tabacism and asbestos exposure are known to be causes of cancer of the respiratory system. The marginal conditional decomposition leading to a recursive system as in model (12) and the associated directed acyclic graph as in fig. (1) will be the following:

$$P_{SES} \cdot P_{A,T|SES} \cdot P_{C|A,T} \tag{12}$$

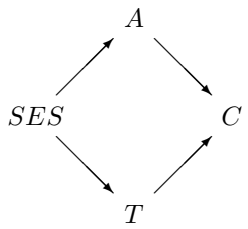

Figure 1: *Socio-economic status, smoking, asbestos exposure and cancer of the respiratory system*

The behavioural mechanism underlying the causal relationship between SES and C would therefore be a differential exposure among socioeconomic groups to tabacism and asbestos. Furthermore, biological mechanisms exist for explaining causal relations between tabacism and asbestos on the one hand, and cancer on the other hand. In this sense SES *is* a cause of C both through a behavioural and a biological mechanism. In other words, structural modelling aims at providing a mechanism—in this case a mixed mechanism (biological and sociological)—explaining a given phenomenon—in this case, different cancer rates among different socio-economic groups (Russo, 2008). Two features particular to model (12) and fig.1 should be stressed. First, this model is not completely recursive because it does not disentangle the process generating A and T conditionally on SES. Reasons for this may be: on the one hand, background knowledge might not provide information whether the

underlying structure has the form $P_{A|SES} \cdot P_{T|A,SES}$ or $P_{T|SES} \cdot P_{A|T,SES}$, on the other hand, the problem of interest focuses on the process generating C and not on the process generating SES, A and T. This is an argument of parsimonious modelling. Second, model (12) and fig.1 incorporate the property $C \perp\!\!\!\perp SES|A,T$ whereas $C \perp\!\!\!\perp SES$ is not true. This is a biological assumption saying that two individuals with the same (A,T) characteristics but with different SES would face a same risk of C although SES and C are not independent (more precisely, $C \perp\!\!\!\perp SES|A,T$ does not imply $C \perp\!\!\!\perp SES$).

Background knowledge plays a major role in the construction of the conceptual model. It is composed of assertions, conjectures and questions, and reflects the present "state of the art". However, background knowledge is only part of the picture as it is made of a vast body of information not always narrowly related to the population of interest. The statistical issue is now to account for data narrowly related to the population of interest in order to make the model built out of background knowledge operational. Then, to establish *generic* causal relations, one must ensure that the structure be stable enough. Stability is required for internal and external validity in the sense of Cook and Campbell (1979).

Structural stability involves two aspects. Firstly, the recursive decomposition should be stable in the sense that each component remains meaningful among different observations. Secondly, the parameters of each conditional distribution should remain numerically invariant. In this context, stability and invariance are relative to a large, and "reasonable", class of interventions and/or of changes of the environment. Thus stability and invariance not only give substance to the concept of structurality but they also are necessary conditions for accumulating statistical information. They are also aimed at defining the population of interest. For example, the recursive system given by model (12) and Figure 1 should be valid *e.g.* among Belgians, when the population of interest is that of Belgium. If this is not so, for example if the decomposition (12) is valid for part of the data only, different models should be developed for the sub-populations.

Unlike purely descriptive approaches, a structural approach is based on the idea that the underlying mechanism can not be discovered by examining only the directly available data and that reference to the observation of other data or knowledge, is indispensable. A structural approach also conveys the idea of distinguishing a basic, or systematic, aspect of the data generating mechanism from an accidental, or non systematic, one. This is operated through the stochastic feature of the statistical model where the random component represents the accidental, or unexplained, aspect of the data generating mechanism whereas the characteristics, or parameters, of these distributions represent the structural aspect. As a corollary, a related problem is to evaluate the extent to which the unexplained part stands for a genuinely random component of the underlying behaviour and/or for the lack of observability of some explanatory factors.

As a summary, the structure of the process generating the observations is, typically, not directly

observable. A structural model is built by merging background knowledge and data, checking for structural stability and invariance in a recursive decomposition. A structural model that takes those three components into account is deemed causal. Namely, each component of the recursive decomposition is taken as an autonomous cause(s)-effect(s) relation. Needless to say, new information, data, or methods can discard the model in favour of another one. In other words, structural models are provisional and deemed causal, to the best of our knowledge.

4.3 Structural modelling need not be based on counterfactuals

In the previous section, we have shown that structural modelling aims at uncovering an underlying data generating process, meaning the mechanisms leading from causes to effects. Structural modelling is therefore apt at both determining the effects of causes and the causes of the effects. Though the treatment/control model is highly effective in testing the effects of causes, structural modelling is not confined to experimental data and may also accommodate for observational data, with or without interventions. As manipulation is not necessarily required, structural modelling may take attributes into account. For example, it can deal with ethnicity or gender. This is at odds with approaches requiring manipulability as, for instance, in Rubin's potential outcome approach or Pearl's causal graphs. Furthermore, it avoids the *possible worlds* trap of counterfactuals.

There are some caveats however. Contrary to the experimental approach, and especially to the double-blind randomised treatment-control model, structural modelling requires a clear assertion of the network of putative relations between causes and effects, *i.e.* a good background knowledge of the problem at hand. It also puts heavy demands on the quality of the data set. In particular, the presence of latent confounding variables, absent from the data set, may play havoc with the results. Finally, the relations between the variables need to be adequately estimated and the estimation method (e.g. covariance structure analysis, multiple equations, linear regression, logit regression, *etc.*) also usually requires a series of assumptions of its own, which have to be met more or less. In addition, the results must be stable to changes of context, in view of reaching "explanatory generalizations", in Woodward and Hitchcock's terms (2003), but this condition is not specific to the structural modelling approach and is needed in general for making causal claims. Indeed, in a Bayesian framework, new data should not lead to posteriors different from priors if the causal explanation is correct. If they do, either the model is not adequate or the context has changed, and both alternatives have to be thoroughly examined.

Though Rubin's counterfactual framework of causal inference suffers from major methodological and epistemological flaws which have been discussed in this paper, we have nevertheless stressed the fact that this does not mean however that counterfactuals should be dropped from the scientific language altogether. We should not throw away the baby with the bath water. As recalled earlier in section 2.2, counterfactual questions are one of several criteria used as aids for causal attribution in

science. These criteria, which include not only counterfactuals but also such issues as the time-order of cause and effect or the absence of alternative explanations, help in building the conceptual model. For example, the counterfactual "If one had taken an aspirin, one's headache would be gone" may suggest taking analgesics into account in the study of headaches, as one knows that headaches may often be relieved by salicylates. On the other hand, the counterfactual "If one had drunk a glass of water, one's headache would be gone" would most probably not lead to incorporating water drinking into the analysis, as the latter usually has no impact on headaches. Counterfactuals can therefore be useful in the process of retaining or rejecting putative causes in/from the structural model.

It is worth emphasising that the structural modelling approach here proposed is not counterfactually based. In particular, the requirement of invariance or stability is not defined in terms of counterfactuals. Various approaches nowadays, for instance Pearl (2000) or Woodward and Hitchcock (2003), provide a *counterfactual* definition of invariance. Such definitions, loosely speaking, define a causal relation invariant if parameters turn out to be numerically stable under intervention, in particular under *hypothetical* interventions—whence their counterfactual character. But this is not suitable to many social science contexts, notably when we analyse observational data (rather than experimental data) or when the putative causes are not manipulable (for instance gender or ethnicity).

Also, the plausibility of the counterfactual approach has also been criticised in the context of policy evaluation. For instance, Reiss (2007) reviews a number of approaches that use counterfactuals to evaluate social policies and shows that they are flawed. However, again, the counterfactual approach ought not to be dismissed altogether, but ought be based on a sound causal analysis.

5 Discussion and conclusion

In this paper, we have followed Pearl's viewpoint (Pearl 2000) that there are presently two approaches to causality: the potential outcome or counterfactual framework as championed most notably by Donald Rubin, and the causal modelling framework *à la* Wright, Haavelmo, Blalock, and others (including Pearl himself). Rubin's potential outcome/counterfactual approach has drawn attention to the important issue of correctly assigning the units to the treatment and to the control groups in non-experimental situations, in order to avoid such pitfalls as the bias resulting from self-selection into the groups. However, this approach suffers from several important methodological and epistemological flaws which jeopardize its use in the social sciences among others. For example, advocating the manipulability of the cause, the counterfactual approach cannot take attributes such as gender or ethnicity into account. Built to examine the effects of causes, the approach is furthermore not adapted to the study of the causes of an effect.

In this paper, in the lines of the causal modelling approach, we developed a general structural

modelling framework based on the following components. First, it relies on a thorough inventory of background knowledge of the issue at hand, composed of assertions, conjectures and questions, and reflecting the present "state of the art" including previous models and studies, scientific theories, expert opinion, etc.. This background knowledge or prior information is required for selecting the *reference population* on the one hand, and for constructing the *conceptual model* composed of the relevant variables and the putative causal relations among them, on the other hand. Two agents may have different background knowledge, or epistemic states as Halpern and Pearl (2005, II) would call them, and could therefore develop two different conceptual models. The latter are always dependent upon the prior information available to the agent. The conceptual model is then transformed into an *operational model* taking the availability of data for the reference population into account. For this purpose, concepts often need to be translated into measurable indicators, the latter having to be valid and reliable. The comparison between the conceptual and the operational model may reveal loss of exogeneity (confounding) due to the fact that some of the variables in the conceptual model are actually not observed and thus become latent.

The operational model is expressed by a *marginal-conditional decomposition*, *i.e.* a decomposition of the multivariate distribution into the product of a sequence of conditional distributions; each term is conditional on an increasing sequence of the components of the random vector and such that each component is congruent with background knowledge. This decomposition may be represented graphically by a *causal* or *Bayesian network* (Halpern and Pearl 2005, I). In particular, if the model is recursive (no feedback), the network becomes a *directed acyclic graph*. However, a model may fail to be completely recursive due, in particular, to incomplete observability. Finally, the model should be *stable*, meaning that the structure of the model should remain similar under changes of contexts, in order to lead to explanatory generalizations.

The marginal-conditional decomposition and the associated graph are non-parametric in the sense that they represent arbitrary functional relations among the variables. Halpern and Pearl (2005, I) then define causality in the language of structural equations, but other modelling approaches can also be used. In other words, the causal network does not imply a particular class of statistical models such as covariance analysis, multiple-equation ordered logistic regression, or whatever. Qualitative methods may also be applied. For example, in the search for the causes of unemployment in a specific reference population, a qualitative method such as case studies based on the analysis of life history narratives can be used to test one's conceptual model. In this situation, the familiar concept of *theoretical saturation* or *informational redundancy* (Sandelowski 1996) corresponds to our concept of stability: new data should eventually not challenge the results of the analysis.

To conclude, contrary to Pearl's or Heckman's causal modelling approaches based on structural equations, the general framework of structural modelling presented here has the advantage of not being linked to a particular quantitative or qualitative approach. It can actually take both into ac-

count. Compared to Rubin's potential outcome/counterfactual framework, the structural modelling framework does not require interventions or manipulation of causes, and can deal both with the effects of causes and with the causes of an effect. It is therefore better suited to observational studies in the social sciences. However, the structural modelling approach requires a thorough background knowledge of the causal network involved. Moreover, a deep understanding of the assignment mechanism is required. Finally, the correspondence between the data and the concepts to be measured should be scrupulously examined. Our approach does not imply that counterfactual questions are irrelevant: they remain useful for setting up the research hypotheses, but not as a basic framework of scientific inquiry.

Acknowledgments 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. F. Russo acknowledges financial support from the FSR-FNRS (Belgium). M. Mouchart acknowledges financial support from the IAP research network Nr P5/24 of the Belgian State (Federal Office for Scientific, Technical and Cultural Affairs). Comments from Wilfredo Quezada are also gratefully acknowledged.

References

Beaglehole R., R. Bonita and T. Kjellström (1994), *Eléments d'épidémiologie*, Organisation mondiale de la Santé, Genève.

Blalock Hubert M. Jr., (1971), The measurement problem: A gap between the languages of theory and research, in Hubert M. Jr. Blalock and Ann Blalock (ed.), *Methodology in social research*, p.5-27. MacGraw-Hill, International Student Edition, London.

Brand and Xie (2007), Identification and estimation of causal effects with time-varying treatments and time-varying outcomes, *Sociological Methodology*, **37**(1), 393-434.

Collins J., Hall N. and Paul L.A. eds, (2004), *Causation and counterfactuals*, The MIT Press, Cambridge, Massachusetts.

Cook, T. D., Campbell D. T. (1979), *Quasi-experimentation. Design and analysis issues for field settings*, Rand MacNally, Chicago.

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

Dawid, A.P. (2007) Counterfactuals, hypotheticals and potential responses: a philosophical examination of statistical causality. In *Causality and Probability in the Sciences*, edited by F. Russo and J. Williamson. College Publications, Texts In Philosophy Series Vol. 5, 503-32, London.

Fienberg S.E. (2006). Comment: Complex causal questions require careful model formulation: discussion of Rubin on experiments with "censoring" due to death, *Statistical Science*, **21**(3), 317-318.

Fisher (1958) *Statistical methods for research workers*, Hafner, New York.

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.

Halpern J.Y. and J. Pearl, (2005, I), Causes and explanations: A structural-model approach, Part I: Causes, *British Journal for the Philosophy of Science*, 56, 843-887.

Halpern J.Y. and J. Pearl (2005, II), Causes and explanations: A structural-model approach, Part II: Explanations, *British Journal for the Philosophy of Science*, 56, 889-911.

Hart H.L.A. and Honoré A.M., (1985), *Causation in the Law*, Clarendon Press, Oxford.

Heckman J., (2005), The Scientific Model of Causality, *Sociological Methodology*, 35(1), 1-97.

Heckman J., (2008), Econometric Causality, IZA DP N° 3525, The Institute for the Study of Labor (IZA), University of Bonn.

Holland, P. (1986) Statistics and causal inference, *Journal of the American Statistical Association*, **81**(396), 945-960.

Hume D. (1748), *An Enquiry Concerning Human Understanding*, Bobbs-Merrill, Indianapolis, 1955.

Lewis D. (1973a), Causation, *Journal of Philosophy*, 70, 556-567. Reprinted with postscripts in Lewis D. (1986), *Philosophical Papers II*, Oxford University Press, Oxford, 159-213.

Lewis D. (1973b), *Counterfactuals*, Harvard University Press, Cambridge.

Lewis D. (1986), Causation, in D. Lewis, *Philosophical Papers Vol. II*, Oxford University Press, Oxford, 159-213.

Lewis D. (2004), Causation as influence, in Collins J., Hall N. and Paul L.A. eds, (2004), *Causation and counterfactuals*, The MIT Press, Cambridge, Massachusetts.

Menzies P. (2001), Counterfactual Theories of Causation, *The Stanford Encyclopedia of Philosophy* (Spring 2001 Edition), Edward N. Zalta (ed.).
URL=http://plato.stanford.edu/archives/spr2001/entries/causation-counterfactual/.

Mill J.S. (1843), *A system of logic*, Longmans, Green and Co., London, 1889 edition.

Morgan S.L. and C. Winship (2007). *Counterfactuals and causal inference*, Cambridge University Press, New York.

Mouchart M., F. Russo and G. Wunsch (2008), Causality, Structural Modeling and Exogoneity, to appear as Chap. 4 in Henriette Engelhardt, Hans-Peter Kohler, Alexia Prskawetz (eds), *Causal Analysis in Population Studies: Concepts, Methods, Applications.* Dordrecht: Springer.

Ni Bhrolchain M. and T. Dyson (2007) On causation in demography: issues and illustrations, *Population and Development Review*, **33**(1), 1-36.

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

Reiss (2007), *Error in Economics: Towards a More Evidence-Based Methodology*, Routledge, London.

RUBIN, D.B. (1974) Estimating causal effects of treatments in randomized and non randomized studies, *Journal of Educational Psychology*, **66**(5), 688-701.

RUBIN, D. (1986) Which ifs have causal answers, *Journal of the American Statistical Association*, **81**(396), 961-962.

RUBIN D.B. (2007), The design versus the analysis of observational studies for causal effects: Parallels with the design of randomized trials, *Statistics in Medicine*, 26(1), 20-36.

RUSSO F.(2008), *Measuring Variations. Causality and Causal Modelling in the Social Sciences*, Methodos Series, Springer, in press.

SANDELOWSKI M. (1996), Sample size in qualitative research, *Research in Nursing & Health*, 18, 170-183.

STALNAKER R. (1968), A Theory of Conditionals. In Nicholas Rescher, ed., *Studies in Logical Theory*, pp. 98-112. American Philosophical Quarterly Monograph Series, 2. Oxford: Blackwell.

WOODWARD J. (2003), *Making Things Happen. A Theory of Causal Explanation*, Oxford University Press, New York.

WOODWARD J. AND HITCHCOCK C. (2003), Explanatory generalizations,Part I: A counterfactual account, *Noûs*, 37(1), 1-24.