DISCUSSION PAPER SERIES

IZA DP No. 354

On the Role of Counterfactuals in Inferring Causal Effects of Treatments

Jochen Kluve

September 2001

Forschungsinstitut
zur Zukunft der Arbeit
Institute for the Study
of Labor

On the Role of Counterfactuals in Inferring Causal Effects of Treatments

Jochen Kluve
Alfred Weber Institute, University of Heidelberg
and IZA, Bonn

Discussion Paper No. 354
September 2001

IZA

P.O. Box 7240
D-53072 Bonn
Germany

Tel.: +49-228-3894-0
Fax: +49-228-3894-210
Email: iza@iza.org

IZA Discussion Paper No. 354
September 2001

ABSTRACT

On the Role of Counterfactuals in Inferring Causal Effects of Treatments [*]

Causal inference in the empirical sciences is based on counterfactuals. This paper presents the counterfactual account of causation in terms of Lewis's possible-world semantics, and reformulates the statistical potential outcome framework and its underlying assumptions using counterfactual conditionals. I discuss varieties of causally meaningful counterfactuals for the case of a finite number of treatments, and illustrate these using a simple set-theoretical framework. The paper proceeds to examine proximity relations between possible worlds, and discusses implications for empirical practice.

JEL Classification: B30, C19, Z00

Keywords: Causation, counterfactuals, possible worlds, treatment effect

Jochen Kluve
Alfred Weber Institute
Grabenstrasse 14
69117 Heidelberg
Germany
Tel.: +49 6221 - 542923
Fax: +49 6221 - 543640
Email: kluve@uni-hd.de

[*] I am grateful to Mark Blaug, Ruth Miquel, Christoph Schmidt, to seminar participants at the Tinbergen Institute Amsterdam and the University of Heidelberg and to participants of an IZA workshop on policy evaluation for valuable comments, to the Cusanuswerk for financial support, and to the Tinbergen Institute Amsterdam for its hospitality. Part of this research was supported by the Deutsche Forschungsgemeinschaft (DFG) under the research grant "SFB 544 – Control of Tropical Infectious Diseases".

In spite of all the evidence that life is discontinuous, a valley of rifts, and that random chance plays a great part in our fates, we go on believing in the continuity of things, in causation and meaning.

– Salman Rushdie, 'the ground beneath her feet' –

1. Introduction

Recent years have seen an increased discussion of causation and varieties of causal models in the fields of econometrics, statistics, computer science, epidemiology, and sociology. Leaving for the moment the century-lasting discourse on accounts of causation in philosophy aside – I will get back to this at a later stage – this increased research on matters of causation in the above-mentioned fields has led to three major approaches to modelling causation currently dominating the debate on causal inference. These are (a) Structural Equation Models (SEM), (b) Potential Outcome Models (POM), and (c) Directed Acyclic Graphs (DAG). In this section, I will first give brief introductions to all three approaches, and then discuss somewhat further how they are perceived in the academic community. Subsequently I will focus on the scope of this paper, its foundations and contributions.

1.1 Modeling Causation: Three Approaches

Structural Equation Models (SEM) as an approach to causation are mainly used in economics and the social sciences. SEM has its origin in path analysis developed by geneticists (Wright 1921, 1934). Founding work in SEM has been done by Haavelmo (1943, 1944) and Koopmans and Hood (1953), work that defined the program of the Cowles Commission and set the stage for modern econometrics (cf. Morgan 1990, Heckman 2000). In fact, SEM has remained the paradigm of causal modeling in contemporary econometrics and the social and behavioral sciences. A set of equations

$$Y = Xb + e$$

is meant to represent a stochastic model in which each equation represents a causal link (Goldberger 1972). All causal connection between Y and X is captured by β, and we infer the causal effect of variation of one element of X relative to its value before variation – holding all other elements of X constant – on Y relative to its value before variation of X. The "all-

other-elements-constant"-clause is well known in economics as the ceteris paribus condition, and in its core goes back to Alfred Marshall (1890 [1965]).

The Potential Outcome Model (POM) of causation predominant in statistics describes a setting in which units are potentially exposed to a set of treatments, and have corresponding outcomes or responses associated with each treatment. The causal connection of interest is the effect on the outcomes of some particular treatment relative to some other particular treatment (often called "control" treatment). Since in reality each unit can only be exposed to one treatment, the other treatment states and associated potential outcomes for the single unit are counterfactuals. In its essence the POM dates back to the work of Neyman (1923 [1990], 1935) and Fisher (1935). Fisher is commonly credited for the invention of randomized experiments, while Neyman was probably the first one to use a model for a treatment effect in which each unit has two responses. Major contributions to the development of the model include Cox (1958), Cochran (1965), and above all Rubin (1974, 1977), who was the first to apply the potential outcome framework to observational studies. See also Rosenbaum (1995a) for further discussion. Due to Rubin's contributions the model is frequently referred to as the "Rubin Model". Related work in economics are models for switching regressions (Quandt 1958, 1972) and the earnings model of Roy (1951). Due to the latter, economic applications occasionally call the POM the "Roy-Rubin-Model".

The use of Directed Acyclical Graphs (DAGs) to assess causal questions is a rather recent phenomenon. The main proponents of graphical approaches to causation are Spirtes, Glymour, and Scheines (2000, first edition 1993) and Pearl (1995, 1998, 2000a).[1] It is difficult to discuss the functioning and mechanisms of DAGs in just a few phrases – for an introduction see the mentioned papers and books. Rather, I want to describe what their advocates think DAGs are aimed at: They are aimed at making causal relations and assumptions and implications in causal models more explicit, in particular more explicit than – in the view of its proponents – other approaches. For instance, Pearl (2000a) claims that recent advances in DAGs have transformed causality from "a concept shrouded in mystery" into a mathematical object with well-defined semantics and well-founded logic. This is another aim of the graphical approach, namely to provide causal talk with a common language helping researchers communicate (Pearl 1995, 1998), an aim that DAGs do not yet live up to in the view of everybody – see the discussion of Pearl (1995), in particular Imbens and Rubin (1995) and Rosenbaum (1995b). Pearl (2000a) strongly emphasizes the gain in

[1] Robins (1986, 1987) offers a graphical approach within the framework of a general counterfactual causal model, related to the POM. See Robins (1995) for how this relates to Pearl's (1995) approach, and see, for

clarity and explicitness gained from causal models based on DAGs in his view. For better or worse, his conclusion is that due to DAGs "causality has been mathematized" (Pearl 2000a).

Naturally, different approaches to questions of causation are viewed differently by proponents of different approaches. For instance, perceptions of SEM as an adequate approach to causation diverge strongly. Pearl (1998) unfolds the idea that the original conceptual strength of SEM along with the clear conception of it among its founding fathers has been lost since, or at least become "obscured". In his perception, social and behavioral scientists – including economists – nowadays struggle for an understanding of either β, or the error term, or both (see Pearl 1998 for examples). In his belief, "the causal content of SEM has gradually escaped the consciousness of SEM practitioners" (Pearl 1998) for two reasons: (i) SEM practitioners have kept causal assumptions implicit in order to gain respectability for SEM, because statisticians, "the arbiters of respectability", abhor assumptions that are not directly testable, and (ii) SEM lacks the notational facility needed to make causal assumptions, as distinct from statistical assumptions, explicit. The latter point means that the SEM founding fathers thought of the equality sign as the asymmetrical relation "is determined by" rather than an algebraic equality, but did not invent a distinct sign for this relation. They were aware of this distinction in meaning, but now their descendants seem to have lost this clear conception – for more on this issue see Pearl (1998), who clearly develops this idea to contrast it with DAGs as a more coherent tool of causal language.

On the other hand, Heckman (2000) is a clear proponent of SEM and forcefully stresses the major role that econometric analysis played in the twentieth century analysis of causal parameters:

> A major contribution of twentieth century econometrics was the recognition that causality and causal parameters are most fruitfully defined within formal economic models and that comparative statics variations within these models formalize the intuition in Marshall's [notion of a ceteris paribus change] and most clearly define causal parameters.

This is how economists define causal effects.[2,3] Heckman (2000) argues that the statistical

---

instance, Greenland (2000), Robins and Greenland (2000), Pearl (2001) as starting points of the literature on causal inference in epidemiology and the health sciences.

[2] It is a bit puzzling, though, that Heckman (2000, p.56) correctly defines causal effects in SEMs as partial derivatives (or as finite differences of some factor holding other factors constant), but previously (p.53) speaks of "marginal causal effects". The derivation given there defines the causal effect. There is no such thing as a marginal causal effect.

[3] A different concept of causation in econometrics is Granger causation, which I will not discuss in this paper. See Granger (1969) for the original account, and, e.g., Holland (1986), Granger (1986), Sobel (1995) for discussion.

POM is simply a version of the econometric causal model.[4] This is in line with his finding that the definition of a causal parameter does not require any statement about what is actually observed or what can be identified from data. A finding that also the SEM founding fathers would have subscribed to, as Pearl (1998) refers to Haavelmo (1943) who explicitly interprets each structural equation as a statement about a hypothetical controlled experiment. Here it becomes clear that the ideas of SEM and POM are not at all so far apart. In fact, a system of structural equations is a system of functions from inputs to potential outcomes (cf. also Greenland 2000). Fortunately, recent years have seen substantial convergence of methods from statistics and econometrics along with increased discourse between the two fields, even though some controversy on who deserves credit for what will most likely remain (cf. Heckman and Hotz 1989, Heckman 1996, Heckman 2000, and Holland 1989, Rubin 1986, 1990, Angrist, Imbens and Rubin 1996).

In summary, one cannot but highly appreciate the vivid debate on causation in the various fields, the expanding amount of causal models suggested, and the many analogies, connections and distinctions that have been drawn between models from different fields. And while one might well wonder whether optimism such as the one recently expressed by Greenland (2000) – "[T]he near future may bring a unified methodology for causal analysis" – seems justified, we do appear to have overcome previous pessimism such as the one expressed by Pearl (1997): "Currently, SEM is used by many and understood by few, while potential-response models are understood by few and used by even fewer." At least, the number of applications has increased substantially; the "understanding" part, though, I would not feel confident to make any judgements about.

1.2 Counterfactuals and Causation

Causation has been a major field of philosophical discussion since at least the pathbreaking works of David Hume (1740a [1992], 1740b [1993], 1748 [1993]), disregarding for the moment early works on causes and effects by the Greek philosophers such as Aristotle and others. Indeed, history has seen an abundance of philosophical approaches to causation, and virtually all of them had repercussions on or counterparts in other scientific fields. The deterministic account of causation of David Hume demanded temporal priority, spatio-temporal contiguity, and constant conjunction as constituent components of a cause-effect relationship. Ideas that fit quite well with Newton-type mechanics. The other way around: Quantum mechanics did force philosophers to re-think possible theories of causation and

[4] The formal equivalence of POM and recursive SEMs has been established by Galles and Pearl (1998).

consider incorporating probabilistic elements (cf. Skyrms 1984). Such a probabilistic account of causation also displays close intertwining between philosophy, statistics, and econometrics (cf. Salmon 1980, Skyrms 1988, Sobel 1995). These incidents of feedback in both directions are manifold between philosophy and other fields.

The volume edited by Sosa and Tooley (1993a) gives an excellent overview of different approaches to causation by various contemporary philosophers, including discussions of the problems inherent to each approach, and to a philosophically structuralist account of causality in general. One of the most remarkable approaches to causation has been the one suggested by David Lewis. Lewis (1973a) developed possible world semantics for counterfactual conditionals, and proceeded to ground his theory of causation on these counterfactuals (Lewis 1973b and 1986).

Causal inference in statistics is based on counterfactuals. In general this view is uncontroversial. A recent approach to causal inference without counterfactuals suggested by Dawid (2000) has met pronounced rejection – cf. the discussion of Dawid (2000), in particular the comments by Pearl (2000b) and Robins and Greenland (2000). The POM has remained the most prominent approach to causal inference in statistics. The previous subsection has already sketched the basic notions along with the historical evolution of the model. The fact that the POM is based on a counterfactual notion of causation has first been pointed out by Glymour (1986) in his discussion of the seminal paper on causal inference in statistics by Holland (1986). However, to my knowledge there has been no further effort to explicitly link the POM and the account of Lewis, even though the mere fact that there indeed is a link has been stated occasionally, and the notion of "closest possible worlds" is not uncommon in talk about causation (cf., e.g., Dawid 2000, Robins and Greenland 2000). The only explicit link I know of is made in Galles and Pearl (1998) who give an axiomatic characterization of causal counterfactuals in comparing logical properties of counterfactuals in structural equation models and Lewis's closest-world semantics.

Why is it that this link has been kept implicit and statistics has sought so little guidance in the philosophical account of counterfactual causation? Some might argue that the POM is a well-defined causal model that does not need further metaphysical or logical underpinning. On the other hand, Pearl (1997) ascribed the – in his viewpoint – perceived failure of the POM to become standard language in statistical inference to it resting "on an esoteric and seemingly metaphysical vocabulary of counterfactual variables that bears no apparent connection to ordinary understanding of cause-effect processes." Given the pervasive number of recent applications, the reluctancy has disappeared. Or has it? In any

case, a fundamental assessment of counterfactuals and their role in causal inference can clarify any researcher's thoughts on causation, the importance of which cannot be overstated (Sobel 1995).

Are counterfactuals "esoteric and metaphysical" entities? And does the idea that they are based on comparative similarity relations between worlds clarify much? Lewis (1973a) replies to the possible objection that these conceptions might be "unclear" with saying that "unclear" is unclear, as "unclear" may express either "ill-understood" or "vague". Counterfactuals and comparative similarity are not ill-understood concepts: they are vague – very vague indeed –, but in a well-understood way (Lewis 1973a).

This paper concentrates on the counterfactual-based nature of the POM. There have been other connections of causes and counterfactuals (Simon and Rescher 1966) and logical theories of counterfactual conditionals (Stalnaker 1984), but – as mentioned above – the outstanding protagonist has been David Lewis with his account of counterfactual logic based on possible-world semantics (Lewis 1973a), a theory on which he then based his theory of causation (Lewis 1973b). Subsequent criticism on some details of his account led him to refine and supplement his original theory, including probabilistic elements, i.e. "chancy counterfactuals" (Lewis 1986).[5] From the perspective of those applying the POM there is need to further investigate its underlying counterfactual nature, and to clarify the counterfactual semantics that the model – implicitly – uses to infer causal relationships.

1.3 Paper Outline

This paper contributes to the literature on causation in explicitly linking the POM and Lewis's account, and in delineating which counterfactual causal questions can be asked, and answered, within the model. The procedure is as follows: I start with looking at how causation is modeled in the empirical sciences and find that there are three major approaches: SEM, POM, and DAG. I pick out the POM as the main causal model of interest in evaluation research. Looking at the model I find that it is formulated in terms of counterfactuals. What, then, are counterfactuals? This leads me to analyze the semantic properties of counterfactuals, and the counterfactual approach to causation in philosophical logic. It turns out that the central element of this approach is the notion of 'possible worlds'. I proceed to connect the counterfactual-based POM with the possible world semantics for counterfactuals, reformulating the POM and its assumptions in terms of counterfactual statements. This

[5] In fact Lewis recently suggested further refinement (Lewis 2000). For the discussion in this paper, however, only the main results from his original theory are relevant. The ongoing discussion is more important from a strictly philosophical point of view (cf. Menzies 2001b).

procedure (i) connects statistical and philosophical understandings of counterfactuals and (ii) adds clarity to the counterfactual nature of the POM. The paper then takes a closer look at this crucial notion of proximity of possible worlds, and finds that within the POM closest possible worlds are defined *a priori*, and merely differ with respect to elements of the treatment set T along with associated outcomes. Therefore, I give a detailed discussion of T using a simple set-theoretical framework. This analysis also elucidates which meaningful counterfactual questions can be asked, and answered.

The remainder is organized as follows. The second section presents the counterfactual account of causation in terms of Lewis's possible-world semantics. It describes the most prominent features of the theory and includes a short assessment of potential metaphysical shortcomings – or, rather, an assessment of general obstacles for theories of causation, and how Lewis's account addresses these. This comprises a concise review of chancy counterfactuals. Section 3 unfolds the POM and its assumptions and reformulates it using the – de facto underlying – ideas of counterfactual conditionals presented in section 2. In this respect, sections 2 and 3 belong together in focussing on foundational aspects of the model.

Sections 4 and 5 slightly change perspective towards a more applied viewpoint. The fourth section delineates how possible counterfactual worlds within the POM differ only with regard to particular treatments and corresponding responses. In general the model allows for finite T, but both theory and practice have focused on only two elements within T, "treatment" and "control". This is intuitively appealing, as a causal effect can only be inferred for one treatment relative to some other treatment. However, as recent results in evaluation research, e.g., have made explicit extensions to multivalued treatment settings in observational studies possible (Imbens 2000, Lechner 2001a), it appears imperative to discuss various issues that arise for causal inference with finite T and relevant counterfactual queries. Section 4 thus presents a set of meaningful counterfactuals, and includes some examples for illustration. A third subsection proceeds to consider notions about proximity relations between possible worlds. I will show that the empirical procedure assumes or constructs closeness ensuring that only the factor we manipulate is different between worlds. If the assumption holds, then closeness is ensured, then the counterfactual conditions hold and the model produces valid inference. The fifth section gives more details on a few specific problems that could arise in practice – this section might be of interest above all for applied social scientists using matching methods. Section 6 concludes.

2. The Counterfactual Account of Causation

Many philosophical discussions of causation begin with – or entail at some stage – the probably most famous quote of David Hume, and present the puzzle which the quote comprises:

> […T]herefore, we may define a cause to be *an object, followed by another, and where all the objects, similar to the first, are followed by subjects similar to the second.* Or in other words, *where, if the first object had not been, the second never had existed.* [Hume 1748 [1993], his italics]

Of course it is not puzzling in the sense that Hume's work represents the foundation of the analysis of causation. His writings – including the major Hume 1740a [1992] – have shaped the examinations of the principles of causation until today. On the other hand, the cited passage is puzzling in the sense that Hume certainly was aware of the regularity-based nature of his first definition, but apparently not of the counterfactual nature of his alternative definition.

Counterfactuals therefore did not play any role in Hume's understanding of causation, and in fact it was not until the 1970s that the link between counterfactuals and causation became object of a thorough philosophical analyses. Before, philosophers had been concerned enough with such a vague concept like counterfactuals themselves – cf. Menzies 2001a for a concise review of early counterfactual theories.

The better understanding of a counterfactual approach to causation was basically due to the development of possible world semantics for counterfactual conditionals by Robert Stalnaker (1984) and David Lewis (1973a, 1986). As the concept of possible worlds plays an important role in this paper, I will review the basic ideas in some detail. The discussion is along the line of thought suggested in Lewis (1973a and 1973b) and further refined in Lewis (1986). "Along the line of thought" in this context means that there are many aspects in the philosophical assessment of causation that are of minor interest to econometricians, statisticians, and social scientists. Philosophical approaches to causation have always tried – or, rather, have always had to try – to give a metaphysical account fundamentally explaining how causation works in our world. Needless to say that there have always been possible objections and counterexamples to each 'structuralist' theory of causation, that the presented theory could not grasp (cf., e.g., Sosa and Tooley 1993b). The empirical sciences however, do not need a causal theory that explains how causal processes work under each and every circumstances in our world. A 'realist' or 'reductionist' approach is sufficient: Clearly, specific

questions like whether it is the table top that causes the table feet to have exactly the length they have, or whether it is rather the table feet causing the table top to assume the spatio-temporal position it occupies, or maybe both, are of subordinate importance. Also, discussions about direction of time in a cause-effect relationship are of minor concern, as a situation in which the effect precedes its cause is extremely unlikely in the empirical sciences.[6]

This paper therefore focuses on those aspects of a counterfactual theory of causation in terms of possible worlds that are of direct use to empirical social scientists. It is both impossible to include a full metaphysical account of this theory, or even to come anywhere near a fair metaphysical account, as well as unnecessary in the given context. I would hope that philosophers would nevertheless agree with (a) the main points I extract from Lewis's theory, and (b) the claim that thinking along these lines can considerably help to sharpen any researcher's thoughts on causal inference based on counterfactuals.

2.1 Possible World Semantics

Lewis's theory of causation employs possible world semantics for counterfactual conditionals, providing truth conditions for counterfactuals in terms of relations between possible worlds. Again, in this exposition we need not worry about the realism of these possible worlds, whether they are "maximally consistent sets of propositions", or "theoretical entities having no independent reality", etc. (cf. Menzies 2001a). Regardless of metaphysical subtlety they provide us with a useful framework of causal thinking.

This section considers the deterministic perspective. Possible world semantics for counterfactuals are based on the main idea of *comparative similarity* between worlds. Given a set of worlds W, according to Lewis (1973b) one world $w_j \in W$ is *closer* to a given world $w_i \in W$ than another world $w_k \in W$ if w_j resembles w_i more than w_k resembles w_i. Naturally, this notion of closeness is based on the idea of w_i being the actual world, and defining $w_j, w_k \in W$ with respect to their proximity to actuality[7]. Lewis imposes two formal constraints on this similarity relation: (i) It produces a weak ordering of worlds such that any two worlds can be ordered with respect to their closeness to the actual world, where "weak" implies that ties

[6] To take an example from economic evaluation of policy interventions: Even if an individual participates in a program because of her expecting the program to raise future earnings, it is not the (potential) future earnings that may have caused her to participate, but the thought (in the present) of the program raising the earnings. Expectation, though directed at the future, is very much a concept of the present.

[7] "Actuality" means the "world of point of view", i.e. "actual" refers at any world w_i to that world w_i itself. Lewis (1973a): "'Actual' is indexical, like 'I' or 'here', or 'now': it depends for its reference on the circumstances of utterance, to wit the world where the utterance is located." The actual world is only one world among others, and we call our world "actual" because it is the one we inhabit, not because it differs in kind from all the rest (cf. Lewis 1973a).

are permitted, but any two worlds are comparable. (ii) The actual world is closest to actuality, resembling itself more than any other world does.

For any two propositions C and E, define the following counterfactual conditionals:

(1a) $C \Box\!\!\rightarrow E$ "If C were (had been) the case, then E would be (have been) the case."[8]

and

(1b) $\sim C \Box\!\!\rightarrow \sim E$ "If C were not (had not been) the case, then E would not be (not have been) the case."

Then the counterfactual conditional $C \Box\!\!\rightarrow E$ is characterized by the following truth condition in terms of the similarity relation:

(2) $C \Box\!\!\rightarrow E$ is true at a world $w_i \in W$ iff either (i) there are no possible C-worlds, or (ii) some C-world where E holds is closer to w_i than is any C-world where E does not hold.

(i) is the trivial case and implies that the counterfactual is vacuously true. From the perspective of w_i being the actual world, the idea of (ii) is that $C \Box\!\!\rightarrow E$ is (nonvacuously) true in the actual world if it takes less of a departure from actuality to make the antecedent true along with its consequent, than it does to make the antecedent true without the consequent (Lewis 1973b). Under the assumption that there must always be one or more closest C-worlds this condition simplifies to $C \Box\!\!\rightarrow E$ being nonvacuously true iff E holds at all the closest C-worlds.

> **Example A.** The classic illustration of Lewis (1973a): "If kangaroos had no tails, they would topple over."

Lewis (1973a) underscores this exemplification in explaining what he thinks that such a counterfactual sentence is supposed to mean: "In any possible state of affairs in which kangaroos have no tails, and which resembles our actual state of affairs as much as kangaroos

[8] $\Box$ represents a *necessity operator* or *'would'-counterfactual*. I omit the consideration of the *possibility operator* or *'might'-counterfactual* $\Diamond$ from the discussion. In terms of truth conditions for counterfactuals, $\Box$ implies truth at all accessible (from the actual world) worlds, while $\Diamond$ implies truth only at some accessible worlds (cf. Lewis 1973a). Causal inference in the social sciences is exclusively interested in the 'would'-counterfactual.

having no tails permits it to, the kangaroos topple over". This statement entails most of what the analysis of counterfactual conditionals is about, namely that a counterfactual sentence corresponds to an actual state of affairs, and that the counterfactual is true if it deviates from actuality only to minimum extent.

> **Example B.** Assuming that it characterizes minimum deviation from actuality, the counterfactual "If John participated in the computer course, he would find a job" is true corresponding to the actual state of affairs in which John does not participate in the computer course and does not find a job.

So far this principal idea considers propositions, not events. Lewis (1973b) extends this setting by pairing the two: To any possible event e there corresponds the proposition O(e) that holds at all and only those worlds where e occurs. Thus, O(e) is the proposition that e occurs, i.e. O(e) is a sentence describing the occurrence of the particular event e, and counterfactual dependence among events is simply counterfactual dependence among the corresponding propositions. We then have a definition of causal dependence:

(3) Let c and e be two distinct possible particular events. Then e causally depends on c iff $O(c) \boxrightarrow O(e)$ and $\sim O(c) \boxrightarrow \sim O(e)$.

This condition states that whether e occurs or not depends on whether c occurs or not. The dependence consists in the truth of the two counterfactuals $O(c) \boxrightarrow O(e)$ and $\sim O(c) \boxrightarrow \sim O(e)$. Consider two cases: first, if c and e do not actually occur, then the second counterfactual is automatically true because its antecedent and consequent are true. Thus, e depends causally on c iff the first counterfactual holds, i.e., iff e would have occurred if c had occurred. Second, if c and e are actual occurrent events, it follows from the second formal condition on the comparative similarity relation (cf. above) that the first counterfactual is automatically true, because the condition implies that a counterfactual with true antecedent and true consequent is itself true. Thus, e depends causally on c iff, if c had not been, e never had occurred. This is exactly Hume's second definition of causation.

To put it simply:

(3a) c causes e iff both c and e are actual occurrent events and if c had not occurred then e would not have occurred.

Or, using the possible world semantics for counterfactuals:

(3b) c causes e iff both O(c) and O(e) are true in the actual world and in the closest (to the actual world) possible world in which O(c) is not true, O(e) is not true.

2.2 Chancy Counterfactuals

What about this counterfactual theory if causation was probabilistic rather then deterministic? It appears natural to compare the counterfactual account of causation – be it from a strictly philosophical viewpoint (Lewis 1973b) or from a statistical perspective (cf. Holland 1986, 1988a, and see below) – with alternative approaches that identify causal dependence in terms of probabilistic relations. There is a rich philosophical literature addressing probabilistic causality in various forms, cf. Reichenbach (1956), Good (1961, 1962), and Suppes (1970) for three classic theories, but also Salmon (1980, 1998) and Suppes (1984), among others.

Probabilistic concepts of causality are used in innumerable contexts of everyday life and science.[9] They have the advantage that they can indeed accommodate many of our everyday experiences, that they appear to make it easy to understand how we can have knowledge of causal relations, in particular in cases where we seem to observe causation, but no determination. As Glymour (1986) puts it: "Technical details aside, causal inference becomes a statistical estimation problem." Could it possibly sound any better to econometricians? But on the other hand, probabilistic concepts of causality have the disadvantage that they do not always coincide appropriately with our intuitive judgements about causal relations, and that causation in terms of percentages may be difficult to conceive.

Nevertheless, contemporary physics – here: quantum mechanics – tells us that our world is full of probabilistic processes that are of causal character (cf. Lewis 1986; or Skyrms 1984 for a discussion of the implications for causality of the Einstein-Podolsky-Rosen (1935) paradox). Thus, Lewis (1986) argues that a theory of causation must accommodate the conceptual possibility of chancy causation. He combines his counterfactual theory of causation with elements of probabilistic concepts, and defines a more general notion of causal dependence in terms of chancy counterfactuals.

(4) $C \mathrel{\Box\!\!\rightarrow} Pr(E)=x$ "If C were the case, then E would be the case with probability x"

The counterfactual is thus an ordinary world counterfactual that can be interpreted according to the semantics above. The *Pr* operator is a probability operator with narrow scope confined

[9] Cf. Salmon (1980) both for an account of the three classic theories as well as a number of illustrative examples. To name but one: "We have strong evidence that exposure to even low levels of radiation can cause leukemia, though only a small percentage of those who are so exposed actually develop leukemia."

to the consequent of the counterfactual. In this context, the definition of causal dependence becomes:

(5) Let c and e be two distinct possible particular events. Then e causally depends on c iff, if c had not occurred, the probability of e's occurring would have been much less than it actually was (given that c occurred).

Obviously this definition comprises the deterministic causal relation in which the probability of the effect along with the cause is 1 and the probability of the effect without the cause is 0. Chancy counterfactuals are thus a straightforward extension to "normal" counterfactuals. For further discussion of probabilistic concepts of causation see, for instance, Hitchcock (1997), Skyrms (1988) and references therein, and Sobel (1995) for the connection to causal inference in the social sciences.

2.3 Applicability

In this subsection I will briefly review some of the problems that arise or need to be addressed in (philosophical) theories of causation. In particular, I will discuss event causation, spurious non-causal dependence, temporal asymmetry, as well as transitivity and preemption.

Above we adopted a definition of causal dependence relating two events. It is, however, not evident that it is events that are the fundamental relata of causal dependence. The philosophical literature clearly distinguishes between event causation and causal theories based on facts or state of affairs (cf. Bennett 1993). Moreover, it implies the necessity to define a certain notion of what is an event – cf. for instance Lewis (1986) for his construction of events as classes of possible spatiotemporal regions. But in general this problem does not arise in the applied social sciences: First, Menzies (2001a) states that even under a metaphysical perspective very different conceptions of events are compatible with the basic definition of causal dependence in terms of counterfactuals. Secondly, it appears straightforward to incorporate the events of interest in empirical research – such as, e.g., a training program, medical treatment, etc. – into the above framework of event causation.

Another element of the above definition requires the causally dependent events to be distinct from each other. This feature rules out what is called spurious non-causal dependence. Consider an example from Kim (1973): Writing the letter "r" twice in succession is a constituent event in the event of writing "Larry". Thus: "If I had not written 'r' twice in succession, I would not have written 'Larry'." The counterfactual is true, but there is no causal

relation between the events. But since the events are not distinct from each other, the relation does not count as causal dependence.

> **Example C.** In econometric evaluations of employment programs one occasionally finds puzzling statements of the sort that the program "increased significantly the employment […] *during* the period of program participation" (Fraker and Maynard 1987, my italics). Clearly, it is impossible to disentangle causal dependence from spurious non-causal dependence if the potential cause (=being employed) is not distinct from the potential effect (=being employed).

What is the temporal structure of causation? As social scientists, both our intuition and the analyses that we deal with in practice suggest clearly that causes would typically precede their effects. But why do we commonly associate causal relations with the temporal direction from past to present or future? Lewis (1979) addresses this point and indeed argues that the direction of causation is the direction of causal dependence, and that it is typically true that events causally depend on earlier events, but not on later events. He notes, however, that the conceptual idea of time-reversed or backward causation cannot be ruled out a priori.

I do not intend to go deep into this analysis, cf. Lewis (1979) and Horwich (1993) to grasp the major issues, but I do want to note the two main points emerging from the discussion: (i) Lewis (1979) defines a determinant for an event as any set of conditions jointly sufficient – given the laws of nature – for the event's occurrence. Looking from the two directions of time, determinants can be causes or *traces* of an event. Any particular fact about a deterministic world is predetermined throughout its past and postdetermined throughout its future. Lewis (1979) observes it to be contingently true that events typically have very few earlier determinants but very many later determinants[10]. This is called asymmetry of overdetermination.

Lewis (1979) combines this de facto temporal asymmetry of causal dependence with (ii) his analysis of the comparative similarity relation between worlds. The comparative similarity analysis implies that the most similar worlds are those in which the actual laws of nature are never violated, and exact similarity regarding particular matters of fact in some spatiotemporal region is an important element of similarity if it can be ensured by a small, local miracle, rather than at the cost of big, global miracle. In connection with the asymmetry of overdetermination, this argument (cf. Lewis 1979, Menzies 2001a for details) implies that it is easier to reconcile a hypothetical change in the actual course of events by preserving the past and allowing for a divergence miracle than by shielding the future from change by virtue

[10] An example (Menzies 2001a): A spherical wave expanding outwards from a point source is a process where each sample of the wave postdetermines what happens at the point at which the wave is emitted.

of a convergence miracle. The main result here is that – given the asymmetry of overdetermination – the present counterfactually depends on the past, but not on the future.

It has to be noted that – strictly speaking – Lewis (1973b) uses the definition given in (3a) and (3b) only as a definition of "causal dependence among actual events". His actual definition of causation is based on the notion of causal chains: Lewis (1973b) states that causal dependence between actual events is sufficient for causation, but not necessary. As it can happen that three actual events c, d, and e are of the form that d would not have occurred without c, and e would not have occurred without d, but e would still have occurred without c, causal dependence may not be transitive. Nonetheless, Lewis (1973b) insists that causation must always be transitive. He therefore extends causal dependence to a transitive relation, where c, d, e, … is a finite sequence of actual particular events such that d causally depends on c, e on d, etc. This he calls causal chain. The definition of causation then becomes

(6) c is a cause of e iff ∃ a causal chain leading from c to e.

This definition ensures transitivity of causation, and it provides a solution to the problem of causal preemption. Causal preemption takes place when the cause of an event preempts something else from causing that event (cf. Horwich 1993 or Menzies 2001b for examples). Using definition (6) it is possible, however, to distinguish preempting actual causes from preempted potential causes.

3. The Potential Outcome Model for Causal Inference: A Reformulation

The statistical model called POM – based mainly on work by Neyman (1923 [1990], 1935), Fisher (1935), Cox (1958), Cochran (1965) Rubin (1974, 1977, 1978, 1980, 1986) – provides a solid ground for causal inference in experimental and observational studies. As it is implicitly cast into a counterfactual framework, it directly relates to – or: is grounded on – many of the aspects of counterfactual logic presented in the previous section. I will give a fairly detailed review of the basic model, and show how it is connected to the possible world semantics presented above. Much of the presentation of the original POM is based on the discussions in Holland (1986, 1988a), since these provide a very clear account of the theory.

3.1 The Causal Model

The logical elements of the POM are a quadruple of the form {U, T, D, Y}. These four elements constitute the primitives of the model. U is a population of N units u $[u_1,...,u_n]$, T is a set of M treatments[11] t $[t_1,...,t_m]$ to which each one of the units u may be exposed, D(u)=t indicates that unit u is actually exposed to a particular treatment t out of T, and Y(u,t) equals the value of the outcome that would be observed if unit $u \in U$ were exposed to treatment $t \in T$. U and T are sets, D is a mapping of U to T, and Y(.) is in general a real-valued function of (u, d).

Note that the response variable Y depends on both the unit u and the treatment t to which the unit is exposed. If u were exposed to some $t_1 \in T$, we would observe the value of the outcome $Y(u,t_1)$, and if u were exposed to some $t_2 \in T$, the observed response value would be $Y(u,t_2)$. The meaning of Y to be a function of pairs (u,t) is that it represents the measurement of some characteristic of u after u has been exposed to $t \in T$. This requirement implies that it must be possible for any unit in U to be potentially exposed to any treatment t out of T. Holland (1988a) emphasizes the importance of this condition: It entails a certain notion of what is a cause, that is of fundamental importance in preventing us from interpreting associational relations as causal ones, like, e.g., associations between sex and income or between race and income. This condition of the POM and its relevance is discussed more extensively in Holland (1986, 1988a, 1988b) and Glymour (1986). The main point that we can derive at this stage is that this condition de facto states that causes must be events.

Call Y the outcome function and let $Y_t(u)=Y(u,t)$. The mapping D is called the assignment rule because it indicates to which treatment each unit is exposed. The observed outcome of each unit $u \in U$ is given by

$$Y_D(u)=Y(u,D(u)),$$

which is the value of Y that is actually observed for unit u. Therefore, the pair $(D(u),Y_D(u))$ – where D(u) indicates the treatment in T to which u is actually exposed – constitutes the observed data for each unit u. Note the distinction between $Y_D(u)$ and $Y_t(u)$: While the former

[11] I chose to stick to the formulation of treatment, rather than, e.g., calling it a "cause" (Holland 1988a) for two reasons: (i) The empirical context of the POM that we are interested in is exactly that of 'treatments' like medicaments in health sciences or policy interventions such as training courses in the social sciences. (ii) From an intuitive linguistic perspective a cause implies an effect. A priori we do not know whether an effect will be observed, the cause of which we desire to infer. So, from the point of view that we do not yet know about an effect and that a zero effect is usually not called an effect, I think that we cannot call T a set of causes a priori. The formulation of treatment is unambiguous.

is the outcome actually observed on unit u, the latter is a potential outcome being actually observed only if D(u)=t.

In the model, treatments are taken as undefined elements of the theory, and effects are defined in terms of these elements (Holland 1988a). The basic causal parameter of interest is

(7) *The unit-level treatment effect (UTE)*:

The unit-level causal effect of treatment $t \in T$ relative to treatment $c \in T$ (as measured by Y) is the difference $Y_t(u)-Y_c(u)=UTE_{tc}(u)$.[12]

There are three important things to note about this definition. First, the causal effect $UTE_{tc}(u)$ is defined at the individual-unit level. Second, $UTE_{tc}(u)$ is the increase in the potential value of $Y_t(u)$ over the potential value of $Y_c(u)$. Third, $UTE_{tc}(u)$ is defined as the causal effect of t relative to c. The following discussion will center around elements number two and three:

Consider $UTE_{tc}(u)$ being the increase in the potential value of $Y_t(u)$, which is what would be observed for the potential outcome if D(u)=t, over the value of $Y_c(u)$, which is what would be observed for the potential outcome if D(u)=c. Here it becomes clear that this is a definition based on causal dependence in counterfactual terms. Define the following set of events:

e_k: Unit $u \in U$ is exposed to treatment $t_k \in T$, i.e. $D(u)=t_k$, and

e^*_k: Unit $u \in U$ has the value $Y_{t_k}(u)$ for variable Y,

where k=1,…,m, so that the number of events for each individual unit is 2×M (as there are N units in U, the total number of events is 2×M×N). Then:

(8) The unit-level causal effect of treatment $t_i \in T$ relative to treatment $t_j \in T$ (as measured by Y) is defined by the difference $Y_{t_i}(u) - Y_{t_j}(u) = \mathrm{UTE}_{t_i t_j}(u)$ iff the counterfactual conditionals $O(e_i) \mathrel{\square\!\!\rightarrow} O(e^*_i)$, $\sim O(e_i) \mathrel{\square\!\!\rightarrow} \sim O(e^*_i)$, and $O(e_j) \mathrel{\square\!\!\rightarrow} O(e^*_j)$, $\sim O(e_j) \mathrel{\square\!\!\rightarrow} \sim O(e^*_j)$ are true.

[12] Note that the two treatments in this definition are denoted with t (like "treatment") and c (like "control"). This already hints at the discussion of randomized assignment of units into an experimental treatment or control group. It also gives a particular flavor to the definition of an effect of one treatment relative to a control treatment, where the term "control" usually implies "no treatment". Moreover, note that the notation E_{tc} is meant to indicate the causal effect of "t relative to c".

To illustrate this reformulation, let us return to the treatment-versus-control case. Define the events:

e_1: Unit u is exposed to treatment t
e_2: Unit u is exposed to treatment c
e^*_1: Unit u has the value $Y_t(u)$ for variable Y
e^*_2: Unit u has the value $Y_c(u)$ for variable Y

In this special case (8) simplifies to:

(8a) The unit-level causal effect of treatment $t \in T$ relative to treatment $c \in T$ (as measured by Y) is defined by the difference $Y_t(u)-Y_c(u)=UTE_{tc}(u)$ iff the counterfactual conditionals $O(e_1) \Box\!\!\rightarrow O(e^*_1)$, $\sim O(e_1) \Box\!\!\rightarrow \sim O(e^*_1)$, and $O(e_2) \Box\!\!\rightarrow O(e^*_2)$, $\sim O(e_2) \Box\!\!\rightarrow \sim O(e^*_2)$ are true.

Recall (3), and note that we have two underlying causal dependencies: e^*_1 causally depends on e_1, and e^*_2 causally depends on e_2. However, to infer either of the two causal dependencies – in this case: that between e_1 and e^*_1 – we need the other one. This is because causal inference can only be made relative to something (cf. below).

Furthermore note the formulation of "distinct possible particular events" in (3), because looking at (8a) and recalling the special case of (3a) we would seem to encounter a problem: Not all of the events e_1, e_2, e^*_1, and e^*_2 can be "actual occurrent events" for a specific unit u, because at the individual-unit level only either e_1 and e^*_1, or e_2 and e^*_2 can be "actually occurrent". Now consider the formulation using possible world semantics for counterfactuals in (3b): In our example, clearly e_1 causes e^*_1, because either both $O(e_1)$ and $O(e^*_1)$ are true in the actual world or, in the closest (to the actual world) possible world, both $O(e_1)$ and $O(e^*_1)$ are not true, because in that closest world $O(e_2)$ and $O(e^*_2)$ are true. It is easy to see that the same argument holds the other way around for e_2 and e^*_2. Note that in this particular case we only have two worlds, and we define the causal effect in one world (the e_1-e^*_1-world) relative to the second and trivially closest world (the e_2- e^*_2-world), whichever one may be the actual world. Thus, also the symmetry of the analysis appears obvious.

This causal analysis can be summarized in four steps: (i) The basic causal parameter of interest is the unit-level treatment effect UTE of some treatment t relative to another treatment c. (ii) UTE is defined iff the pairs of events e_1 and e^*_1, and e_2 and e^*_2 (as defined above) are

both causally dependent. (iii) These pairs of events are causally dependent iff the counterfactual conditionals $O(e_1) \Box\!\!\rightarrow O(e^*_1)$, $\sim O(e_1) \Box\!\!\rightarrow \sim O(e^*_1)$, and $O(e_2) \Box\!\!\rightarrow O(e^*_2)$, $\sim O(e_2) \Box\!\!\rightarrow \sim O(e^*_2)$ are true. (iv) A counterfactual conditional of the type $O(a) \Box\!\!\rightarrow O(b)$ is true in the actual world iff any a-world along with b is closer to actuality than any a-world without b.

Return to (7) and the third aspect we noted, that $UTE_{tc}(u)$ is defined as the causal effect of t relative to c. Indeed, the effect of one treatment is always relative to the effect of another treatment. We can only draw inference on the cause of an effect by relating two effects of two distinct causes (or: potential causes, or treatments). This relativity condition is one of the central aspects of the POM: a causal relation between a treatment and an effect can be identified from "measuring" two alternative states of an outcome variable given some unit has been exposed to some treatment or another. As Glymour (1986) puts it: "Causation is a relation between two treatments and two possible variable states. The notion of t causing Y_t, without specification of any alternative treatment, or any alternative state of Y, is not defined." Glymour (1986) regards this as an improvement on the bare counterfactual account of causal relations, and he presents an example supporting his argument. I include this example here, because I think it is highly elucidating with respect to the relativity condition and the idea of possible worlds behind it [my italics]:

> **Example D.** My Uncle Schlomo smoked two packs of cigarettes a day, and I am firmly convinced that smoking two packs of cigarettes a day caused him to get lung cancer. But it may not be true that in the closest possible world in which Uncle Schlomo did not smoke two packs a day, he did not contract cancer. Reflecting on Schlomo's addictive personality, and his general weakness of will, it may well be that the closest possible world in which Schlomo did not smoke two packs of cigarettes a day is a world in which he smoked three packs a day. I can reconcile this reflection with the counterfactual analysis of causality by supposing […] that 'smoking two packs of cigarettes a day caused him to get lung cancer' is elliptical speech, and what is meant, but not said, is that smoking two packs of cigarettes a day, *rather than not smoking at all*, caused Schlomo to contract lung cancer.

This example highlights many of the inherent features of the model. I want to point out two more aspects that will require further discussion below. First, many of our casual causal thoughts are based on just that idea of inferring causal relations from saying "doing a relative to not-doing-a", where not-doing-a may equal doing-nothing, and for most analyses this is exactly the causal question of interest. Second, this example nicely raises the question of how can we identify the closest world to actuality (or, the world that we are interested in and use as a base category), or, for the very least, how can we derive any notion what the features of this closest world are supposed to look like. I will examine these two issues in section 4.

Let us rephrase Example D using the framework introduced above. This will accentuate the way the model works, even though it does not exactly correspond to an "exposure to treatment" context and may thus sound a bit odd at first sight, and even though it disregards issues of timing, disturbing factors etc. Define the following events:

e_1: Schlomo ("Unit u") smokes 2 packs of cigarettes a day. ("Treatment t")
e_2: Schlomo does not smoke at all. ("Treatment c")
e^*_1: Schlomo contracts lung cancer. ("$Y_t(u)$")
e^*_2: Schlomo does not contract lung cancer. ("$Y_c(u)$")

The outcome variable Y can be regarded as "health status" or something similar. According to (3) and (3a,b) e_1 causes e^*_1 because both are actual occurrent events and if the former had not occurred then the latter would not have occurred. This alternative world is specified by e_2 and e^*_2. The two crucial things about this example are (i) that we have an explicit specification of the closest possible world to actuality, and (ii) that this closest world to actuality is defined by the fact that the actual occurrent events do not occur, i.e. $e_2 = \sim e_1$, and $e^*_2 = \sim e^*_1$.

Feature (i) is far from unusual, because in fact the relativity condition in (7) *per definitionem* specifies the closest world – i.e. the "treatment-c-world" – to the actual world – i.e. the "treatment-t-world".[13] Causal inference about treatment t is based on the counterfactual relation to what would have happened under exposure to treatment c. In that sense the model does not depend on *searching* for the closest possible world, but rather on *justifying* the choice of what is claimed to be the closest possible world, or the relevant alternative world. Feature (ii) of our example says that in the relevant alternative world $e_2 = \sim e_1$, and $e^*_2 = \sim e^*_1$, i.e. treatment c is merely the absence of treatment t. This is a special case in which UTE is defined under the simplified condition that only $O(e_1) \boxrightarrow O(e_3)$ and $\sim O(e_1) \boxrightarrow \sim O(e_3)$ need to be true.[14] In fact, this is how causal inference is usually made, and it is what Glymour (1986) means with "elliptical speech": We infer the causal effect of something relative to not-that-something. The general definition (8) accommodates this simplified case, but above all it accommodates the case of "something relative to something else", i.e. in the example a valid causal relation between treatment t and its outcome (= the e_1-

[13] As pointed out in footnote 7 "actual world" means something like "world of departure", or "world of point of view", or "world of interest" due to the analysis being symmetric: If the c-world is closest to the t-world, then also the t-world is closest to the c-world, and the inferred causal effects are the same in magnitude, but in opposite direction or with opposite sign.

[14] Because $O(e_2) \boxrightarrow O(e^*_2) = O(\sim e_1) \boxrightarrow O(\sim e^*_1) = \sim O(e_1) \boxrightarrow \sim O(e^*_1)$, and $\sim O(e_2) \boxrightarrow \sim O(e^*_2) = \sim O(\sim e_1) \boxrightarrow \sim O(\sim e^*_1) = O(e_1) \boxrightarrow O(e^*_1)$.

e^*_1-world) and some alternative treatment c and its outcome (= the e_2-e^*_2-world). It is important to note, however, that according to (8) this causal relation between the e_1-e^*_1-world and the e_2-e^*_1-world does entail some statements about the non-occurrence of either event, as both conditionals $\sim O(e_1) \Box\!\!\rightarrow \sim O(e_3)$ and $\sim O(e_2) \Box\!\!\rightarrow \sim O(e_4)$ need to be true, and therefore some consideration of the no-treatment-state is at least implicit.

3.2 Applicability

I have emphasized before that the definition of causal dependence is based on the notion of distinct possible particular events. And it is with respect to any two distinct treatments t and c that we face the fundamental problem of causal inference in practice: It is impossible to simultaneously observe $Y_t(u)$ and $Y_c(u)$, and therefore also the causal effect $UTE_{tc}(u)$ is never directly observable. This is why we need counterfactual statements about possible worlds. The counterfactual statement enters in the form illustrated in Examples A and B: We have an actual state of affairs – i.e. for instance we observe u being exposed to t and responding with $Y_t(u)$. We then infer the causal effect of t on Y(u) by relating this actual state of affairs to the counterfactual statement about how u would have responded – i.e. $Y_c(u)$ – if u had been exposed to c, where the counterfactual characterizes a possible world with minimum deviation from actuality.

Holland (1988a) stresses how the POM makes the unobservability of the causal effect explicit in separating the observed pair (D,Y_D) from the function Y. In fact, a model for causal inference can be interpreted as some specification of the values of Y. In Holland's (1988a) words, causal inference consists of combining (a) a causal model or causal theory, (b) assumptions about data collection, and (c) observed data to draw conclusions about causal parameters. The causal model has been laid out above – this section focuses on how and under what assumptions this model can be applied. I will first discuss one basic assumption and subsequently review the conditions under which we can identify the causal effect from data. Usually this implies imposing restrictions on either Y and/or U that make it possible to assess two potential outcomes for a single unit and therefore infer meaningful causal statements.

The *stable-unit-treatment-value-assumption* (SUTVA) is the pivotal assumption ensuring that the causal framework of the POM is adequate in practice. SUTVA is advocated by Rubin (1980, 1986) to play a key role in deciding which questions are formulated well enough to have causal answers. It is the a priori assumption that the value of Y for unit u when exposed to treatment t is the same independent of (i) the mechanism that is used to

assign t to u, and (ii) what treatments d the other units v≠u receive, and that this holds for all n units within U and m treatments within T. SUTVA is violated when, for instance, there is interference between units that leads to different outcomes depending on the treatment other units received – i.e. Y_{tu} depends on whether v≠u received t or some other d ∈ T – or there exist unrepresented versions of treatment or versions of treatments leading to "technical errors" (Neyman 1935)[15] – i.e. Y_{tu} depends on which (uninteded) version of treatment t unit u was exposed to.

In the counterfactual conditionals framework SUTVA can be represented as follows. Define the following set of events:

e_{ij}: Unit $u_i \in U$ is exposed to treatment $t_j \in T$, i.e. $D(u_i)=t_j$, and

e^*_{ij}: Unit $u_i \in U$ has the value $Y_{t_j}(u_i)$ for variable Y,

where i=1,…,n and j=1,…,m, so that the number of events is 2×N×M. Then SUTVA assumes that the counterfactual conditionals $O(e_{ij}) \Box\!\!\rightarrow O(e^*_{ij})$ and $\sim O(e_{ij}) \Box\!\!\rightarrow \sim O(e^*_{ij})$ are true $\forall$ e_{ij} and e^*_{ij} with i=1,…,n and j=1,…,m independent of (i) the mechanism leading events e_{ij} to occur, and (ii) the other occurring events e_{kl}, k=1,…,n, l=1,…m, i≠k, j≠l.[16]

The fundamental requirements of SUTVA therefore appear to go hand in hand with the more philosophical underpinnings I have presented above. To conclude with Rubin (1986, his italics):

[15] For further detail cf. Rubin (1980) and Rubin's (1990) discussion of Neyman (1923 [1990]). Many aspects of the POM for causal inference (in particular the notion of potential outcomes) are already present in the work of Neyman (cf. also Speed 1990), where they are based on the methodological discussion of agricultural experiments. In that context, possible violations of SUTVA are apparent: How should one avoid neighboring plots treated differently (by, e.g., different fertilizers) to "interfere" given nature's powers (wind, rain etc.), or how can one claim that each bag of fertilizer represents exactly the same treatment as any other bag of fertilizer (cf. Rubin 1986)? Moreover, as Rubin (1990) points out, interference between units can be a major issue when studying medical treatments for infectious diseases, or educational treatments given to children who interact with each other.

[16] I use individual statements of the form $O(e_{ij}) \; o\!\!\rightarrow O(e^*_{ij})$ to represent the POM using Lewis's semantics. Galles and Pearl (1998) use an equivalent representation that could be called a "covering counterfactual" and translate Lewis's statement $A \Box\!\!\rightarrow B$ as "If we force a set of variables to have the values A, a second set of variables will have the values B". If A stands for a set of values $x_1,\ldots,x_n$ of the variables $X_1,\ldots,X_n$ and B for a set of values $y_1,\ldots,y_m$ of $Y_1,\ldots,Y_m$ then

$$A \Box\!\!\rightarrow B \equiv \begin{array}{l} Y^1_{x_1\ldots x_n}(u) = y_1 \,\& \\ Y^2_{x_1\ldots x_n}(u) = y_2 \,\& \\ \ldots \\ Y^m_{x_1\ldots x_n}(u) = y_n \end{array}$$

> [T]he crucial point [...] is that we are not ready to estimate, test, or even logically discuss *causal effects* until units, treatments, and outcomes have been defined in such a way that SUTVA is plausible.

Unit homogeneity is a name given by Holland (1986a) to the assumption that the responses of all units to a particular treatment are the same, i.e. that units respond homogeneously to each treatment:

$$Y_t(u)=Y_t(v) \qquad \forall\, u,v \in U, \text{ and all } t \in T.$$

This is a partial specification of Y in that it restricts the values that Y can take on but does not specify them completely. The assumption is only likely to be justified if one can claim to be working with a homogeneous sample. Under the assumption of unit homogeneity, the causal effect of a treatment t relative to a treatment c is given by

$$UTE_{tc}(u)=Y_t(u) - Y_c(v)=Y_t(v)-Y_c(u)$$

for any two distinct units u and v in U. In this case, UTE_{tc} is a constant and does not depend on the unit under scrutiny. Evidently, unit homogeneity solves the fundamental problem of causal inference in that we only need to measure the two (observable) outcomes $Y_{D(u)=t}(u)$ and $Y_{D(v)=c}(v)$ for two units u and v to infer the causal effect of treatment t relative to treatment c on *any* unit within U. This assumption affects condition (8) simply by providing us with an easy answer with respect to what happens for any unit of U in the closest possible world to that very unit. In the closest possible world to the one in which a unit u is exposed to t and responds with $Y_t(u)$, u would be exposed to c and respond with $Y_c(u)$ by assumption, and under unit homogeneity the latter value of Y is given by the response of any other unit $v \neq u$ of U being (or having been) exposed to c.

Unless unit homogeneity holds, individual effects are impossible to observe. Therefore, one of the most important causal parameters of interest is the average causal effect of a treatment, as it represents a useful summary of the unit-level treatment effects[17]. Let E(.) denote the average value of the argument.

[17] In practice, further questions arise as to whether it is e.g. the "average treatment effect on the treated", or the "average treatment effect on the population" that is the causal parameter of interest. Cf. Heckman (1992) and Heckman, LaLonde, and Smith (1999) for discussion, and Angrist, Imbens, and Rubin (1996a) for more on the POM and identification of the "local average treatment effect" (LATE).

(9) The *average treatment effect (ATE)* of treatment $t \in T$ relative to treatment $c \in T$ is the expected value of the unit-level difference $Y_t(u)-Y_c(u)$ over all $u \in U$, i.e. $ATE_{tc}=E(UTE_{tc})=E(Y_t-Y_c)=E(Y_t)-E(Y_c)$.

The ATE is an unobserved quantity, since expectations of Y for both t and c are taken over the full range of U. In practice it is only possible to observe D(u) and $Y_D(u)$ over U, and therefore only the joint distribution of D and Y_D rather than D and $\{Y_t: t \in T\}$. The average value of the observed outcome Y_D among all those units actually exposed to a particular treatment $t \in T$ can be written as $E(Y_D|D=t)$. For the two particular treatments t and c this becomes $E(Y_D|D=t)=E(Y_t|D=t)$ and $E(Y_D|D=c)=E(Y_c|D=c)$, respectively. These two quantities are always observed in the data, and we can therefore define:

(10) The *prima facie average treatment effect (FATE)*[18] of a treatment $t \in T$ relative to a treatment $c \in T$ is the difference in average responses between those units actually exposed to t and those units actually exposed to c, i.e. $FATE_{tc}=E(Y_t|D=t)-E(Y_c|D=c)$.

The distinction between FATE and ATE emphasizes the fact that the quantity that we can always compute from the data (FATE) does in general not equal the quantity about which we desire to draw inferences (ATE). This results from the difference between $E(Y_t)$ and $E(Y_c)$ on the one hand and $E(Y_t|D=t)$ and $E(Y_c|D=c)$ on the other hand. The former are averages of Y over all of U and constitute ATE, while the latter are averages of Y over only those units in U actually exposed to t and c, respectively, and constitute FATE.

The two quantities are only equal when *independence* holds. Suppose that the determination of which treatment a unit is exposed to is statistically independent of all other variables, in particular the response function. Following – as common practice – Dawid's (1979) notation of independence using the symbol "$\perp\!\!\!\perp$", this can be written as $D \perp\!\!\!\perp \{Y_t: t \in T\}$. Then $E(Y_t|D=t)=E(Y_t)$ for any $t \in T$, and we have:

[18] This follows Holland (1988a) who calls this parameter "prima facie average causal effect FACE". It is not to be confused with a "prima facie cause" as defined by Suppes (1970) in his probabilistic theory of causation (cf. Suppes (1970) for the original definition and e.g. Sobel (1995) or Salmon (1980) for a discussion): Given two time values t and t* with $t<t^*$, the event c_t is a prima facie cause of the event e_{t^*} if $Prob(e_{t^*}|c_t)>Prob(e_{t^*})$, i.e. c temporally precedes e and is positively relevant to it. As Holland (1986) points out, the association between cause and effect defining a prima facie cause is indeed a causal effect under "certain conditions that have wide use in science", while on the other hand FACE "is not always a causal effect". This is also why I prefer the labeling FATE to FACE.

(11) If $D \perp\!\!\!\perp \{Y_t : t \in T\}$, then the *prima facie average treatment effect* of a treatment $t \in T$ relative to a treatment $c \in T$ is equal to the *average treatment effect* of t relative to c, i.e. $FATE_{tc}=E(Y_t|D=t)-E(Y_c|D=c)=E(Y_t)-E(Y_c)=ATE_{tc}$.

The independence assumption is the key point to the applicability of the model, as it allows us to draw inferences on the unobserved causal parameter of interest, the ATE, directly from the FATE, which we can always compute or estimate from the data.

Under which conditions is independence likely to hold? The most probable case we have in practice is that of a randomized experiment, in which – coarsely speaking – units are randomly assigned to different treatments, so that the initial population and the subpopulations in the treatments do not differ from each other *on average*. This makes (11) likely to hold, thus yielding the ATE from the FATE. Holland (1988a) describes the relation between randomization and independence as follows: Independence is an assumption about the data collection process, i.e. about the relation of D and Y over the population U, while randomization is a physical process that gives plausibility to the independence assumption in many important cases. For instance, if U were infinite, then the law of large numbers together with randomization would imply that (almost) every realization of D would be independent of $\{Y_t\}$. However, randomization does not necessarily make independence plausible in each and every case, as randomization does not assure that each and every experiment is "adequately mixed", but only that "adequate mixing" is probable (Leamer 1983). To take the simplest example, imagine that U consisted only of very few units. Then the plain physical act of randomization would not render the independence assumption plausible.

What does it mean when we talk about populations that do not "differ" from each other, and "adequate mixing" in randomized experiments? This becomes clear when we introduce other variables into the model. So far Y was the only variable measured on the units u – apart from the treatment indicator D. Let us now add a variable X to the model, where X can be real-valued or vector-valued. In principle, X(u,t) is defined on $U \times T$ and depends on both u and t. However, there is a special class of X-variables that are of specific interest, as defined in Holland (1988a):

(12) X is a *covariate* if X(u,t) does not depend on t for any $u \in U$.

Holland initially calls this class of variables "attributes" (in Holland 1986), but converts to (12) as the preferable definition because it corresponds to the usual experimental usage. If we

consider specifically the values the X-variables take on for units *prior to treatment*, then the X-variables are always covariates.[19] For a post-treatment concomitant, however, the possibility that X(u,t) does depend on t cannot be excluded and "must be decided" (Holland 1988a), and if this the case, then X is not a covariate in the sense of (12).[20] Randomization *on average* guarantees balancing of covariates – observable and unobservable – across subpopulations in different treatments, which in turn makes the independence assumption plausible, so that (11) holds and we can infer the ATE from the FATE. In the words of Rosenbaum and Rubin (1983): With "properly collected data in a randomized trial", X is known to include *all* covariates that are both used to assign treatments and possibly related to the response $\{Y_t\}$.

The introduction of covariates into the model becomes even more important in cases in which we do not have randomization and therefore cannot arrange the values of D(u) to achieve independence. In such an *observational study* we are still interested in inferring causal effects of treatments, but now – differing from a randomized setting – D is not automatically independent of $\{Y_t\}$. Given a (n observable) covariate (vector of covariates) X one could check the distribution of X for subgroups in each treatment by comparing the values of Prob(X=x|D=t) across the values of $t \in T$ (Holland 1988a). If there is evidence that Prob(X=x|D=t) depends on t, then the independence assumption may not appear plausible in the observational study. Instead, in the non-experimental setting one usually builds on a weaker conditional independence assumption which says that treatment assignment and the response are conditionally independent given a vector of covariates:

(13) [Rosenbaum and Rubin 1983:] Treatment assigment is *strongly ignorable* if the response $\{Y_t : t \in T\}$ is conditionally independent of treatment assignment D given the observed covariates X, i.e. $\{Y_t\} \perp\!\!\!\perp D|X$, and $0<Prob(D=t|X)<1$.

Rosenbaum and Rubin (1983) show that (13) also holds for a balancing score B(X) defined as a function of the observed covariates X such that the conditional distribution of X given B(X)

[19] Note that this does not exclude unobservables a priori. In fact, it is difficult to express this feature. Holland (1988a) speaks of "variables measured on units prior to […] treatment" always being covariates, but I find that misleading, in a sense that a priori the definition of a covariate says nothing about whether the variable can be observed or not, and "measurement" implies observation. The point is: A pre-treatment variable is always a covariate, be it observable or not.

[20] Cf. Rosenbaum (1984) for a discussion of adjustments for a concomitant variable that has been affected by treatment.

is the same for the exposure groups (D=t), i.e. $X \perp\!\!\!\perp D|B(X)$.[21] Rosenbaum and Rubin identify all functions of X that are balancing scores; the most trivial one being B(X)=X, and the coarsest one being the *propensity score*: The propensity score $Prob(D=t|X)=P_t(X)$ is of particular interest in practice, as it reduces the potential problem of conditioning on a high-dimensional X – if X is vector-valued – to conditioning on a scalar, provided that $P_t(X)$ is known.

Strong ignorability is the basis for all causal inference on covariate-adjusted treatment effects in observational studies (Holland 1988a). Adjusting for covariates yields the *covariate-adjusted prima facie average treatment effect*, or C-FATE[22], based on conditional expectations:

(14) $C\text{-}FATE_{tc}=E\{E(Y_t|D=t, X) – E(Y_c|D=c, X)\}$.

Just like the FATE, the C-FATE does in general not equal the desired ATE. This only holds under conditional independence:

$$\begin{aligned} C\text{-}FATE_{tc} &= E\{E(Y_t|D=t, X) – E(Y_c|D=c, X)\} \\ &= E\{E(Y_t|X) – E(Y_c|X)\} \\ &= E(Y_t)– E(Y_c) \\ &= ATE_{tc} \end{aligned}$$

This finding concludes the discussion of the POM for causal inference, as we have now discussed all relevant features of the theory as well as the circumstances under which the model can be applied in randomized trials and observational studies. For further discussion cf. Rubin (1974, 1977, 1986), Holland (1986, 1988a), Holland and Rubin (1983, 1988), Rosenbaum (1984, 1995a), Rosenbaum and Rubin (1983, 1984a, 1984b), and Angrist, Imbens, and Rubin (1996a).

[21] In fact Rosenbaum and Rubin (1983) prove this property for the two-treatment case D={0,1}. The extension of this result to multivalued treatments is shown in Imbens (2000), and Lechner (2001a). The main result, however, is that of Rosenbaum and Rubin.
[22] "C-FACE" in Holland (1988a).

4. Comparing Possible Worlds

The program of causal inference is clear from the previous sections: to draw inference about the effect of some treatment t on some response variable Y. It is therefore necessary to establish a counterfactual state of the world – i.e. some other possible world – characterized by an alternative treatment c (with respective associated outcome) to which we can causally relate the treatment-t-world. We have seen that c could be either simply not-t or any other – distinct possible particular – treatment. In this section I will show what such possible worlds look like in the POM, and give some guidelines on the choice of appropriate alternative worlds for inferring and interpreting causal relations.

From the discussion above, in particular the definition of causation based on possible world relations, one could infer that something like "the quest for the closest possible world" is at the heart of the problem of causal inference in statistics. But this is not the case – at least not in a sense that we would have to compare multitudes of worlds and judge degrees of proximity between them. In fact, the POM simply defines closest possible worlds. Section 4.1 considers these ex ante defined worlds and examines relations between them with respect to meaningful causal interpretations. Subsequently, section 4.2 gives an account of proximity relations between worlds and discusses why we might not always be interested in the "closest" possible world, and under what circumstances this can be problematic.

4.1 Varieties of Counterfactuals

By definition the closest world to actuality is the one to which we relate the causal comparison: If we want to infer the causal effect of treatment t relative to treatment c, then we set the c-world as the closest world to the t-world. Were we instead interested in the effect of t relative to some other treatment c*, then we would establish the c*-world as closest to the t-world. This is clear from (8) and the explanation I have given in section 3. The idea that we can consider these worlds as being "close" to each other becomes clear in the experimental context: The t- and the c-world are based on the same background – ideally represented by a large number of concurrent covariates – and they merely differ in the fact that in the t-world units are exposed to treatment t, while in the c-world units are exposed to c.

Let us slightly refine the discussion. First, call T_1 the world solely defined by treatment t_1 *and* associated outcome Y_{t_1}, T_2 the world defined by t_2 and Y_{t_2} etc., so that the M treatments $t_1,\ldots,t_m$ with associated outcomes $Y_{t_1},\ldots,Y_{t_m}$ constitute the M subsets $T_1,\ldots,T_m$

within the set of treatment-worlds T. Second, define the causal effect of some treatment t_i relative to another treatment t_j as

$$\Delta_{t_i t_j} = Y_{t_i} - Y_{t_j} ,$$

disregarding whether we are looking at unit-level or average treatment effects.

Let W denote the universal set comprising all possible worlds that differ only with respect to the characteristic "treatment and associated outcome", so that $T \subseteq W$.[23] In general, T does not need to equal W, if we regard T as comprising just those worlds where we can either control the types of treatment t_i or at least observe them, i.e. T is meant to comprise those worlds with well-defined types of treatment. The complement T' of T is then given from $W=T \cup T'$ and contains treatment-worlds we can neither control nor observe. For both groups, however, it is in principle possible to construct valid comparisons, and thus infer causal relations, as T' can always be defined recursively as "everything that is not T". The relationship is depicted in the Venn diagram in Figure 1a, where W is represented by the rectangle. In Figure 1a, $T=T_1 \cup T_2 \cup \ldots \cup T_m$ and $T_i \cap T_j = \varnothing$ for all $T_i, T_j \in T$, i.e. T is meant to consist of exactly M mutually exclusive treatment-worlds. Clearly, once more this captures the idea of distinct possible particular treatments – Each T_i is well-defined, there is no interference between the T_i worlds, and no unrepresented versions of treatment exist. The special case for $T'=\varnothing$, and thus W=T, is displayed in Figure 1b.

Let us consider the complement T' and what is meant by "everything that is not T". One could argue that "not T" is a well-defined treatment and should be included as one subset in T, yielding T=W. The distinction, however, illustrates the difference between what could be called a *controlled control treatment* and an *uncontrolled comparison treatment*.

Consider the case in which treatment can only take on two values, M=2, the classic treatment-t-versus-control-c setting, where the causal effect of interest is that of t relative to c. In a randomized medical trial, where t is the medicament under study and c is a placebo, c represents a controlled control treatment. It is (a) controlled by the experimenter, and (b) a distinct alternative treatment in its own right, which is not merely characterized by the absence of t, i.e. $T_c \neq T_t'$. Therefore, in this case, $W=\{T,T'\}$ with $T=\{T_t,T_c\}$ and $T'=(T_t \cup T_c)'$, where T' is some unspecified treatment outside T characterized by not given the medicament

[23] This account is still in the framework of the POM of section 3, and thus considers a finite number of treatments. For an extension of the model to the case where the set of treatments is not finite see Pratt and Schlaifer (1988).

and not given the placebo. Of course, T' might not be of interest in the study, or we might not even be able to obtain any information about it, but nonetheless it is an implicit part of the study.

< Figure 1 about here >

On the other hand, consider the case of an observational study in labor economics, for instance, aiming at evaluating some government training program (=treatment t). In this case, the "alternative treatment" c is characterized retrospectively by the absence of training, so that c represents an uncontrolled comparison treatment. It is (a) not under control of the researcher, and (b) not defined on its own, but just by the absence of t, i.e. $T_c=T_t'$. Therefore, W=T with $T=\{T_t,T_c\}$ or, equivalently, $W=\{T,T'\}$ with $T=T_t$ and $T'=T_c$.

Note that the distinction between "controlled control treatment" and "uncontrolled comparison treatment" is about the distinction itself, and does not imply that the one can be used for valid causal inference, and the other cannot. But it is important to note the difference. Clearly placebos are used to learn something about "not given the medication", and in that respect they may perform even better than "actually not given the medication", because with placebos the control units cannot be influenced by knowing that they are not given the medication. Use of placebos ensures that the response is to the treatment itself, not the idea of treatment. Hence, the controlled control treatment gives a well-defined alternative to t, while the uncontrolled comparison treatment necessarily remains more vague.[24] However, we will see that this need not be a disadvantage in interpreting results. It has to be noted, though, that if the control treatment is not well-specified, and the treatment shows no effect relative to the control treatment, then it might well be that the control treatment is or contains a pre-empted potential cause (cf. section 2.3) of the same effect, i.e. both treatment and control have the same causal effect on the response variable, which the causal comparison between the two cannot reveal.

In the next step, let us adopt the notion of treatment c meaning the absence of any treatment, be it controlled or uncontrolled, defined uniquely or recursively. Thus, denote $T_c=T_0$ and let $T=T_0,T_1,\ldots T_{m-1}$ with M-1 "real" treatments and the "null" treatment, W=T.

[24] Experimental settings do not necessarily imply a well-defined null treatment. While this is possible in medical experimental studies of the type described above, it is far more difficult in experimental studies in labor economics, e.g., due to the length of treatment (several months of participation in a training program) and the difficulty of defining a proper alternative (cf. later this section). One example is the experimental evaluation of the National Supported Work Demonstration (NSW) in the US: "Those assigned to the treatment group received all the benefits of the NSW program, while those assigned to the control group *were left to fend for themselves*." (LaLonde 1986, my italics)

Figure 1c depicts the case for M=2 and $T_t=T_1$, $T_0=T_1'$. As T_1 is the world with treatment t_1, and T_0 the world with treatment t_0:

$$(15) \qquad \Delta_{t_1 t_1'} = Y_{t_1} - Y_{t_1'} = Y_{t_1} - Y_{t_0} = \Delta_{t_1 t_0}$$

This is the basic case which almost all causal inference studies are based on. We have just two treatment-worlds differing only by the treatment under study, where the alternative world is characterized by the null treatment which equals the absence of treatment. This setting is intuitively appealing: The closest-world relation is obvious, and the interpretation of results is straightforward.

Let us extend this to the case where M>2, i.e. we have at least two "real" treatments besides the null treatment. Figure 1d illustrates the simplest case for M=3 under consideration and $T_0=(T_i \cup T_j)'$. The case of multivalued treatment has several important implications for interpretation. First, consider particular treatments t_i, t_j, t_k and the following decomposition:

$$(16) \qquad \begin{aligned} \Delta_{t_i t_j} &= Y_{t_i} - Y_{t_j} \\ &= Y_{t_i} - Y_{t_k} - Y_{t_j} + Y_{t_k} \\ &= (Y_{t_i} - Y_{t_k}) - (Y_{t_j} - Y_{t_k}) \\ &= \Delta_{t_i t_k} - \Delta_{t_j t_k} \end{aligned}$$

Of particular interest is the special case where $t_k=t_0$.

$$(16a) \qquad \Delta_{t_i t_j} = \Delta_{t_i t_0} - \Delta_{t_j t_0}$$

(Of course, if $t_k \neq t_0$ in (16) we can only use the decomposition if M>3). Expressions (16) and (16a) nicely show that any causal comparison between two treatments is implicitly always related to any other baseline-treatment within T. The case in which the null treatment is the baseline (16a) is of particular interest, since we have seen above that we are usually used to inferring causal effects relative to the null treatment. This relating of causal comparisons between two treatments – neither of which is the null – to the null treatment is also necessary to identify the level of effects.

For the M=2 case property (15) has shown that the causal comparison of some treatment t_i relative to the absence of t_i equals the comparison of t_i to the null. Unfortunately,

this convenient feature does not hold for a causal comparison of t_i relative to t_i' in the case of M>2. There are two aspects to the t_i-versus-t_i' relation in this context. First, we have the basic result that

(17) $\Delta_{t_i t_i'} \neq \Delta_{t_i t_0}$

because $t_i' \neq t_0$ and $Y_{t_i'} \neq Y_{t_0}$. This can be seen when we consider what the effect of t_i relative to t_i' actually is:

$$\begin{aligned} \Delta_{t_i t_i'} &= Y_{t_i} - Y_{t_i'} \\ &= Y_{t_i} - F(Y_{t_0}, Y_{t_1}, ..., Y_{t_{i-1}}, Y_{t_{i+1}}, ..., Y_{t_{m-1}}) \\ &= Y_{t_i} - \sum_{\substack{k=0 \\ k \neq i}}^{M-1} w_k Y_{t_k} \end{aligned} \tag{18}$$

where $\sum w_k = 1$ and, for instance,

(18a) $w_k = \overline{w} = \dfrac{1}{M-1}$ or (18b) $w_k = \dfrac{P(t=t_k)}{\sum_{r=0, r\neq i}^{M-1} P(t=t_r)} = \dfrac{P(t=t_k)}{1-P(t=t_i)}$.

The causal effect of treatment t_i relative to t_i' as given in (18) is therefore the difference in outcomes under t_i and t_i' (first line), which equals the difference between the outcome under t_i and some function of the outcomes under all other treatments except t_i (second line), which could in an empirical application equal the difference between the outcome under t_i and the weighted sum of all other outcomes (third line). I will refer to the function of the outcomes under all other treatments in T_i' as the *absolute counterfactual* to treatment t_i, as it is a summary expression of all counterfactual possible worlds. Examples of weight functions for empirical work are given as (18a) equal weights, and (18b) the probability of exposure to a particular program (that is not t_i) relative to the sum of probabilities of exposure to any program that is not t_i.[25]

The second aspect to the t_i-versus-t_i' relation is that the complements to particular treatments cannot be used as a common baseline, i.e.

(19) $\Delta_{t_i t_j} \neq \Delta_{t_i t_i'} - \Delta_{t_j t_j'}$

because clearly

$$\begin{aligned}\Delta_{t_i t_i'} - \Delta_{t_j t_j'} &= (Y_{t_i} - \sum_{\substack{k=0 \\ k\neq i}}^{M-1} w_k Y_{t_k}) - (Y_{t_j} - \sum_{\substack{l=0 \\ l\neq j}}^{M-1} v_l Y_{t_l}) \\ &= Y_{t_i} - Y_{t_j} - \sum_{\substack{k=0 \\ k\neq i}}^{M-1} w_k Y_{t_k} + \sum_{\substack{l=0 \\ l\neq j}}^{M-1} v_l Y_{t_l} \\ &= \Delta_{t_i t_j} - (\sum_{\substack{k=0 \\ k\neq i}}^{M-1} w_k Y_{t_k} - \sum_{\substack{l=0 \\ l\neq j}}^{M-1} v_l Y_{t_l})\end{aligned}$$

Table 1 presents an overview of different causal queries and the corresponding counterfactuals.

< Table 1 about here>

In the M=2 case, there are two possibilities, either (a) $t_0=t_1'$ or (b) $t_0\neq t_1'$. The first case (a) is the usual one, and applies for observational studies. The second case (b) comprises two possibilities depending on a relevance criterion. On the one hand, if $t_0\neq t_1'$, so that there exists a T' world besides $T=\{T_0,T_1\}$, and t_1' is considered *irrelevant* for some reason, such as t_0 being explicitly specified – like in an experimental study –, then this implies that T' is irrelevant. On the other hand, if one has reason to believe that $t_0\neq t_1'$ and if T' is *relevant*, then there are two further possibilities: Either (i) one has some usable information about T', then this converts to the M>2 case, or (ii) one does not have such information, which hints at a violation of SUTVA because there exist unrepresented versions of treatment. Usually (as in the agricultural setting of Neyman 1923 [1990]) one thinks of unrepresented versions of treatment as unrepresented versions of the "actual" treatment – in this case, however, T' comprises unrepresented versions of the null treatment.

For the M>2 case, as we have a variety of well-defined treatments, it makes sense to assume that we have a specific t_0 (even if it is defined via the absence of all other treatments) and thus W=T. Table 1 depicts some possible counterfactual comparisons. First, the causal

[25] This expression has been used, e.g., in Lechner (2001b). See also section 5.

effect of a particular treatment could be inferred relative to the null treatment. As in the M=2 case, this would be the causal question of interest in most cases. Second, one could construct the causal comparison of a particular treatment relative to any other treatment within T. In interpreting the effect it should then (a) be pointed out why this is considered to be a causal question of interest, and (b) be noted that any other treatment (besides the two we relate) can be used as baseline. The most relevant baseline is the null treatment, and in fact it should be considered in any case in order to identify the level of the inferred effect.[26] The third possible counterfactual for M>2 in Table 1 relates a specific treatment t_i to a function of the outcomes of all other treatments except t_i. This I labeled absolute counterfactual. It infers the causal effect of some treatment relative to (an appropriate combination of) all other alternative treatments. This could be a weighted average as given in (18). In a sense this is similar to the t_1'=*anything-that-is-not-*t_1-case for M=2, with the decisive difference that now it is "everything", not "anything", expressing the fact that all alternative treatments are well-defined – and that the corresponding outcomes can therefore be appropriately weighted in an empirical study. With respect to the absolute counterfactual, it can be of particular interest to compare the null to the summary over all other treatments to infer whether the introduction of the overall set of treatments yielded any positive response.

Finally, it should be noted that one could of course construct many more counterfactuals. For instance, one could use causal relations between treatments as a baseline for causal relations between other treatments, or construct the comparison between a particular treatment and a weighted combination of some, but not all of the alternative treatments, etc. That, however, is pure mechanics, and I suppose it will be difficult – though not impossible – to unfold the exact causal interpretations of such counterfactuals.

4.2 Illustration

This short subsection entails a few examples that further illustrate some of the ideas unfolded in the previous section, and shows why we need a clear conception of the T worlds for causal inference.

> **Example E.** In the M=2 case, why can it can be insightful to distinguish a known or well-defined no-treatment state ($t_0 \neq t_1'$) from a no-treatment state defined merely by the absence of treatment ($t_0 = t_1'$)? Imagine a researcher plans to evaluate some government training program for the economically disadvantaged in a nonexperimental setting.

[26] If the effect of t_i is positive relative to the null, and the effect of t_j is negative relative to the null, then the effect of t_i is strongly positive relative to t_j. Looking only at the last effect does not reveal the negative effect of t_j relative to the null. Similarly, the effect of t_i relative to t_j could be positive, but still the effects of both of them could be negative relative to the null.

She constructs some retrospective comparison group defined by not having participated in the program. However, training usually takes time. Assume an average of 2 months in this example. What did comparison group units do during that time? Remain unemployed, continue job search, do nothing, take private training course, etc.? Maybe some of that, maybe all of that, maybe none of that. In most cases, the data doesn't tell. Thus, as it is impossible to open this black box, one needs to make some assumption about the comparison treatment. It is then fairly convenient to define the no-treatment state as just that, the absence of the treatment under study. The causal effect is that of the training program relative to any other possible (but unobserved) alternative action the program participants would have engaged in had they not participated. Clearly, this is quite different from the explicit specification of the no-treatment state in an experimental medical study (t_0=placebo).

Example F. Consider the problem of compliance. For instance, in a long-term medical study one could in principle distinguish four groups: those assigned to treatment who are good compliers, those assigned to treatment who are poor compliers, those assigned placebos who are good compliers, and those assigned placebos who are poor compliers. In principle this defines four different treatments, and only the randomized comparison gives the correct inference. Cf. Rosenbaum (1995b) for a discussion, and Angrist, Imbens and Rubin (1996a) for more on compliance.

Example G. An observational study by Larsson (2000) evaluates labour market programs in Sweden. In the study M=3, and the treatments are Youth Practice, Labor Market Training, Non-participation (=Null). In personal communication with the author the interpretation was given that the null treatment comprises a state of job search rather than non-participation. This finding has several implications: (a) If one has usable information to distinguish job searchers from non-participants, this converts to a case of M=4 with treatments YP, LMT, job search, non-participation (=Null). (b) If in fact all individuals in the null treatment are in job search, this changes the counterfactual question, and the causal inference is on the effect of YP (or LMT) relative to job search, and not relative to non-participation. (c) If the null treatment comprises both individuals in job search and non-participants, this hints at a violation of SUTVA.

4.3 Comparative Similarity

Much of what has been said in the previous sections referred to notions of possible worlds, to entities that exist parallel to – or in addition, or besides – something we referred to as actuality, and we viewed these possible worlds in terms of similarity or comparability. Although I am convinced that the intuitive conceivability of this concept of actuality and "surrounding" possible worlds is straightforward, I will discuss a few inherent aspects in this section. More about the foundations of this concept can be found in Lewis (1973a, Ch4), in which, for instance, he replies to a fictitious questioner asking him what sort of thing possible worlds are [Lewis' italics, my underscoring]:

> I can only ask him [the questioner] to admit that he knows what sort of thing our actual world is, and then explain that other worlds are more things of *that* sort, <u>differing not</u>

> in kind but only in what goes on at them. Our actual world is only one world among others. We call it alone actual not because it differs in kind from all the rest but because it is the world we inhabit.[27]

The POM picks up this idea in that the treatment worlds $T_i \in T$ do certainly not differ in kind from each other, but only in the treatment by which they are defined. In fact, the treatment worlds T_i are defined to differ from each other in exactly two aspects: the treatment t_i that "goes on" at each (distinct possible particular) treatment world T_i, and the outcome Y_{t_i} associated with t_i on world T_i. Even though the treatment worlds T_i coexist, they only represent potentialities: Recall that treatments are defined on units u, so that we actually have the two differing features $t_i(u)$ and $Y_{t_i}(u)$ on each $T_i(u)$. However, for each u only one particular $T_i(u)$ is realized. This realized $T_i(u)$ represents actuality. From the point of view of actuality, the other treatment worlds are entities that might be called "ways things could have been" (Lewis 1973a). These he calls *possible worlds*.

In applying the POM we do not search for possible worlds. The treatment worlds T_i are assumed to be, and defined to be the possible worlds. And the treatment world $T_j \neq T_i$ to which we relate treatment world T_i is defined to be the closest possible world to infer the causal effect of treatment t_i relative to treatment t_j. Let us examine this a bit further and return to the example with Clark Glymour's uncle Schlomo from section 3.

> **Example D [ctd].** In the actual world Schlomo smokes 2 packs of cigarettes a day. It has been argued that the closest possible world to that actual world might be the one in which Schlomo smokes 3 packs day. Nonetheless, we choose to define the world in which he does not smoke at all as the closest world. Therefore we infer a causal comparison of actuality with Schlomo smoking 2 packs a day – associated with the outcome "contracting lung cancer" – relative to the closest possible world in which he does not smoke at all – associated with the outcome "not contracting lung cancer". At first sight this appears to be an easy solution. But note that in this causally relating an actual world to a closest world by definition it is implicitly assumed that *no other element* than the one we either manipulate (in an experiment) or study (in an observational study) and the outcome associated with this element differs. This is just the *ceteris paribus* clause of economics. In the example, the element that differs is the reduction in packs of cigarettes a day from two to zero. But in the zero-world it may be that (a) Schlomo takes the healthy road and stops drinking and starts working out a lot, or that he takes on even worse compensatory vices instead, and e.g. starts taking cocaine. Moreover, underlying changes in the zero-world may be that (b) the quitting makes Schlomo lazy, silent, unmotivated, or maybe vivid, lively, energetic instead. Elements (a) refer to observable differences, and elements (b) to unobservable, and the examples show that both (a) and (b) could point into either a positive or negative direction in health terms. For a causal comparison of the actual world relative to the defined closest possible world it is necessary to either control for these potential

[27] Cf. also section 2.1 and footnote 7.

> changes or to ensure that the assumption that these differences equal zero appears credible. Clearly, in the example with Schlomo neither seems very likely.[28]

This is where the proximity relation enters: The 3-pack-world may be closer to the 2-pack-world in a sense that it is more likely that all other factors are the same. But still it may not be the alternative world that we are interested in for inferring a causal relation. Therefore, the causal effect (on some outcome) of some treatment t_i relative to some alternative treatment t_j is based on T_j being (i) the counterfactual world of interest relative to T_i and (ii) the closest possible world by definition. Closeness has to be ensured either by assuming proximity, i.e. on plausibility grounds, or by controlling for background factors establishing that they are the same in both T_i and T_j, in particular those that could potentially influence the outcome.

Finally, one could proceed to discuss distance measures between possible worlds. Thinking of closest possible worlds, this discussion comes up naturally: If M>2, then there is the actual world and at least two alternative worlds. Now which one of the alternative worlds is closer to actuality? This line of thought is a bit misleading, because actually – as shown above – we would condition on background factors (or plausibility grounds, if these can be conditioned on) to ensure that the possible worlds are equidistant. If T_i and T_j differ only in elements t_i/t_j and Y_{t_i} / Y_{t_j}, and T_i and T_k differ only in elements t_i/t_k and Y_{t_i} / Y_{t_k}, then this is also true for T_j and T_k, and the three worlds are pairwise equidistant. In practice, this argument would hold in an experiment with randomized controlled exposure to a set of treatments. In an observational study, however, differences between groups of units across treatment worlds do arise. Therefore, distances between treatment worlds can in principle be measured by appropriately weighting the background factors, or calculating weight functions such as (18b) based on propensity scores for each treatment world (cf. also section 5). I will not pursue this discussion any further here, but conclude with Lewis' observation on the constructability of such proximity measures (Lewis 1973a, his italics):

> We could, however, define exact distance measures […] for […] constructions of ersatz worlds. At worst, we might need a few numerical parameters. For instance, we might define one similarity measure for distribution of matter and another for distribution of fields, and we would then need to choose a weighting parameter to tell us how to combine these in arriving at the overall similarity of two worlds. All this would be easy work for those who like that sort of thing, and would yield an exact

[28] The predominant difficulty in this example is the duration of treatment (=smoking a certain number of packs of cigarettes a day), which goes on for decades. How could you possibly control for background factors, observable or not, over such a long time period? – In any case, I think that this example is easier to reconcile with chancy counterfactuals: Given that Schlomo had not smoked 2 packs of cigarettes a day, the probability of his contracting lung cancer would have been (much) lower.

measure of *something* – something that we might be tempted to regard as the similarity 'distance' between worlds.

Clearly, this is about distance measures between any two worlds. In the POM we depart from actuality and look for that alternative treatment world that sets all these differences to zero, except for the treatment and its associated outcome.

5. Practical Considerations

The following section briefly discusses some specific problems that might arise in empirical work, in particular in observational studies. For causal contrasts, in the POM it is often assumed that the N units are exposed to the M treatments at equal shares. In practice this is unlikely to hold, even in a randomized experiment. And while in a randomized experiment this does not necessarily influence causal comparisons between two treatments – because subsamples are still balanced –, it would indeed affect the absolute counterfactual: Participation probabilities are no longer the same, and therefore the assumption of equal weights is unrealistic. Still, participation probabilities in a randomized experiment would be known: This problem is more severe in an observational study when the null treatment group as a comparison group is unknown and has to be constructed. This usually implies having to estimate the participation probabilities.

The absolute counterfactual of (18) as a summary measure for the effect of some treatment relative to all other treatments can also be represented using a weighted aggregate of the pairwise causal comparisons between the particular treatment and all other treatments:

$$(20) \qquad \Delta_{t_i t_i'} = Y_{t_i} - \sum_{\substack{k=0 \\ k \neq i}}^{M-1} w_k Y_{t_k} = \sum_{\substack{k=0 \\ k \neq i}}^{M-1} w_k (Y_{t_i} - Y_{t_k}) = \sum_{\substack{k=0 \\ k \neq i}}^{M-1} w_k \Delta_{t_i t_k}$$

This expression retains the causal interpretation of the effect of treatment t_i relative to the hypothetical state of random exposure to any other program that is not t_i. Lechner (2001b) uses this expression and calls it the *composite treatment effect*. Furthermore, Lechner (2001b) shows that in an applied observational study it does indeed make a difference whether one assesses t_i' as $T'=T_0$ using a binary probability model or t_i' as $T'=\{T_0 \cup T_1 \cup ... \cup T_{i-1} \cup T_{i+1} \cup ... \cup T_{m-1}\}$ using a multinomial probability model (and then equations (18) or (20)).

The first results in an insufficient specification of the alternative state by aggregating groups into one alternative group without taking into account the different composition of subgroups, while the second appears to correctly disentangle the desired absolute counterfactual. This finding emphasizes the importance attributed to the definition of T' in section 4.1.

Note, though, that this is a problem arising *in practice*. In theory – or in an ideal randomized experiment – the calculation of the absolute counterfactual in the multinomial case would equal the T versus $T'=T_0$ in the binomial case, as it captures all relevant alternatives to a particular T. This holds even if participation probabilities differ across subgroups. In an applied observational study, however, this does not hold, because group compositions do differ, because binomial and multinomial probability models would yield different participation probability estimates, and because the multinomial case compares treatments pairwise (and the overall equivalence above would require a common support of covariates over all subsamples). This is unlikely to be achieved in an observational study, also because in practice heterogeneous programs are aimed at heterogeneous groups.

Finally it has to be noted that even though these two causal comparisons should be the same in theory but differ in practice, they can nonetheless be calculated. However, a meaningful causal interpretation may be difficult to derive (cf. also Lechner 2001b). This again points to the fact that it may not be a problem to mechanically produce various causal comparisons in mechanically ensuring proximity between worlds, but that it may well be a problem to give these comparisons a clear-cut causal interpretation.

6. Conclusion

This paper has tried to add clarity to the understanding, applying and interpreting of the potential outcome model for causal inference commonly used in statistics and econometrics. At the outset we have found that there are three predominant approaches at modeling causation in the empirical sciences: SEM, POM, and DAG. Being the model of particular interest in evaluation research, the paper has focused on the POM and its inherent counterfactual nature. In order to clarify what is actually meant by counterfactual statements, this paper has presented the main elements of the counterfactual account of causation in terms of Lewis's possible-world semantics. This included the basic notions of counterfactual logic, and some of the problems associated with philosophical approaches to causation in general, and the counterfactual approach in particular. I have pointed out that the pivotal notion of

Lewis's account is that of "closest possible worlds".

The paper has then proceeded to explicitly reformulate the potential outcome model for causal inference using counterfactual conditionals. The main ingredients of the POM – such as SUTVA – have a straightforward and elucidating representation in terms of counterfactual events and their truth conditions. I have discussed various causally meaningful counterfactuals that arise in applications of the potential outcome model with a finite number of treatments, and illustrated these using a simple set-theoretical framework. The main result in this respect is that the notion of closeness and proximity between possible worlds is an inherent part of the statistical model, yet one that is implicitly used and taken care of. Causal comparisons in the POM *a priori* assume that possible worlds differ only with respect to the particular treatment and the associated response. However, one has to be aware of the fact that mechanical productions of proximity do not necessarily generate clear-cut causal statements.

References

Angrist, Joshua D., Guido W. Imbens, and Donald B. Rubin (1996a), "Identification of Causal Effects Using Instrumental Variables" (with discussion), *Journal of the American Statistical Association* 91, 444-472.

Angrist, Joshua D., Guido W. Imbens, and Donald B. Rubin (1996b), "Rejoinder", *Journal of the American Statistical Association* 91, 468-472.

Bennett, Jonathan (1993), "Event Causation: The Counterfactual Analysis", in E. Sosa and M. Tooley (eds), *Causation*, Oxford: Oxford University Press.

Cochran, W.G. (1965), "The Planning of Observational Studies of Human Populations", (with discussion) *Journal of the Royal Statistical Society* Series A 128, 234-266.

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

Dawid, A.P. (1979), "Conditional Independence in Statistical Theory", *Journal of the Royal Statistical Society* Series B 41, 1-31.

Dawid, A.P. (2000), "Causal Inference Without Counterfactuals", (with discussion) *Journal of the American Statistical Association* 95, 407-448.

Einstein, A., B. Podolsky, and N. Rosen (1935), "Can the Quantum Mechanical Description of Reality Be Considered Complete?", *Physical Review* 47, 777-780.

Fisher, Ronald A. (1935), *The Design of Experiments*, Oliver & Boyd: Edinburgh.

Fraker, Thomas, and Rebecca Maynard (1987), "The Adequacy of Comparison Group Designs for Evaluations of Employment-Related Programs", *Journal of Human Resources* 22, 194-227.

Galles, David and Judea Pearl (1998), "An Axiomatic Characterization of Causal Counterfactuals", *Foundations of Science* 3, 151-182.

Good, I.J. (1961, 1962, 1963), "A Causal Calculus I and II", *British Journal for the Philosophy of Science* 44, 305-318, and 45, 43-51, and "Errata and Corrigenda", ibid. 46, 88.

Glymour, Clark (1986), "Statistics and Metaphysics", comment on Holland (1986), *Journal of the American Statistical Association* 81, 964-966.

Goldberger, Arthur (1972), "Structural Equation Methods in the Social Sciences", *Econometrica* 40, 979-1001.

Granger, Clive (1969), "Investigating Causal Relations by Econometric Models and Cross-Spectral Methods", *Econometrica* 37, 424-438.

Granger, Clive (1986), "Comment on 'Statistics and Causal Inference' by P.W. Holland", *Journal of the American Statistical Association* 81, 967-968.

Greenland, Sander (2000), "Causal Analysis in the Health Sciences", *Journal of the American Statistical Association* 95, 286-289.

Haavelmo, Trygve (1943), "The Statistical Implications of a System of Simultaneous Equations", *Econometrica* 11, 1-12.

Haavelmo, Trygve (1944), "The Probability Approach in Econometrics", *Econometrica* 12, supplement.

Heckman, James J. (1992), "Randomization and Social Policy Evaluation", in C. Manski and I. Garfinkel (eds), *Evaluating Welfare and Training Programs*, Harvard University Press: Cambridge, MA, 201-230.

Heckman James J. (1996), "Comment on 'Identification of Causal Effects Using Instrumental Variables'" by J.D. Angrist, G.W. Imbens, and D.B. Rubin, *Journal of the American Statistical Association* 91, 459-462.

Heckman, James J. (2000), "Causal Parameters and Policy Analysis in Economics: A Twentieth Century Retrospective", *Quarterly Journal of Economics* 115, 45-97.

Heckman, James J., and V. Joseph Hotz (1989), "Rejoinder to Comments on 'Choosing Among Alternative Nonexperimental Methods for Estimating the Impact of Social Programs: The Case of Manpower Training'", *Journal of the American Statistical Association* 84, 878-880.

Heckman, James J., Robert J. LaLonde, and Jeffrey A. Smith (1999), "The Economics and Econometrics of Active Labor Market Programs", in O. Ashenfelter and D. Card (eds.), *Handbook of Labor Economics*, Vol. III, Ch. 31, Amsterdam: North-Holland.

Hitchcock, Christopher (1997), "Probabilistic Causation", *The Stanford Encyclopedia of Philosophy* (Spring 2001 Edition), Edward N. Zalta (ed.), URL= http://plato.stanford.edu/entries/causation~probabilistic/

Holland, Paul W. (1986), "Statistics and Causal Inference", (with discussion), *Journal of the American Statistical Association* 81, 945-970.

Holland, Paul W. (1988a), "Causal Inference, Path Analysis, and Recursive Structural Equation Models", *Sociological Methodology* 18, 449-484.

Holland, Paul W. (1988b), "Causal Mechanism or Causal Effect: Which Is Best for Statistical Science?", Comment on "Employment Discrimination and Statistical Science" by A. P. Dempster, Statistical Science 3, 186-188.

Holland, Paul W. (1989), "It's Very Clear": Comment on 'Choosing Among Alternative Nonexperimental Methods for Estimating the Impact of Social Programs: The Case of Manpower Training' by J.J. Heckman and V.J. Hotz, *Journal of the American Statistical Association* 84, 875-877.

Holland, Paul W. and Donald B. Rubin (1983), "On Lord's Paradox", in H. Wainer and S.Messick (eds), *Principals of Modern Psychological Measurement*, L. Erlbaum, Hillsdale, NJ.

Holland, Paul W. and Donald B. Rubin (1988), "Causal Inference in Retrospective Studies", *Evaluation Review* 12, 203-231.

Horwich, Paul (1993), "Lewis's Programme", in E. Sosa and M. Tooley (eds), *Causation*, Oxford: Oxford University Press.

Hume, David (1740a [1992]), *Treatise of Human Nature*, as reprinted by Prometheus Books, Buffalo, NY.

Hume David (1740b [1993]), "An Abstract of A Treatise of Human Nature", in *An Enquiry Concerning Human Understanding*, as reprinted by Hackett Publishing Co., Indianapolis, IN, Eric Steinberg (ed.).

Hume, David (1748 [1993]), *An Enquiry Concerning Human Understanding*, as reprinted by Hackett Publishing Co., Indianapolis, IN, Eric Steinberg (ed.), based on the 1777 posthumous edition.

Imbens, Guido W. (2000), "The Role of Propensity Score in Estimating Dose-Response Functions", *Biometrika* 87, 706-710.

Imbens, Guido W., and Donald B. Rubin (1995), "Discussion of 'Causal Diagrams for empirical research' by J. Pearl, *Biometrika* 82, 694-695.

Koopmans, Tjalling and William Hood (1953), "The Estimation of Simultaneous Linear Economic Relationships", in W. Hood and T. Koopmans (eds), *Studies in Econometric Method*, Chapman & Hall: New York.

LaLonde, Robert J. (1986), "Evaluating the Econometric Evaluations of Training Programs with Experimental Data", *American Economic Review* 76, 604-620.

Larsson, Laura (2000), "Evaluation of Swedish youth labour market programmes", Uppsala University, Dept. of Economics *Working paper* 2000-6.

Leamer, Edward (1983), "Let's Take the Con Out of Econometrics", *American Economic Review* 73, 31-43.

Lechner, Michael (2001a), "Identification and Estimation of Causal Effects of Multiple Treatments under the Conditional Independence Assumption", in Lechner, M. and F. Pfeiffer (eds.), *Econometric Evaluation of Labour Market Policies*, forthcoming.

Lechner, Michael (2001b), "Program Heterogeneity and Propensity Score Matching: An Application to the Evaluation of Active Labor Market Policies", *Review of Economics and Statistics*, forthcoming.

Lewis, David (1973a), *Counterfactuals*, Oxford: Blackwell.

Lewis, David (1973b), "Causation", *Journal of Philosophy* 70, 556-567.

Lewis, David (1979), "Counterfactual Dependence and Time's Arrow", *NOÛS* 13, 455-476.

Lewis, David (1986), *Philosophical Papers: Volume II*, Oxford: Oxford University Press.

Lewis, David (2000), "Causation as Influence", *Journal of Philosophy* 98, 182-197.

Marshall, Alfred (1890 [1965]), *Principles of Economics*, 8th ed., repr., Macmillan: London.

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

Menzies, Peter (2001b), "Difference-Making in Context", in J. Collins, N. Hall, and L. Paul (eds), *Causation and Counterfactuals*, MIT Press, forthcoming.

Morgan, Mary (1990), *The History of Econometric Ideas*, Cambridge University Press: Cambridge.

Neyman, Jerzy (1923 [1990]), "On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9.", translated and edited by D.M. Sabrowska and T.P. Speed from the Polish original, which appeared in Roczniki Nauk Rolniczych Tom X (1923), 1-51 (Annals of Agriculture), *Statistical Science* 5, 465-472.

Neyman, Jerzy (1935), with co-operation by K. Iwaszkiewicz, and S. Kolodziejczyk, "Statistical Problems in Agricultural Experimentation", (with discussion), *Supplement to the Journal of the Royal Statistical Society* 2, 107-180.

Pearl, Judea (1995), "Causal diagrams for empirical research", (with discussion), *Biometrika* 82, 669-710.

Pearl, Judea (1997), "The New Challenge: From a Century of Statistics to an Age of Causation", *Compting Science and Statistics* 29, 415-423.

Pearl, Judea (1998), "Graphs, Causality, and Structural Equation Models", *Sociological Methods and Research* 27, 226-284.

Pearl, Judea (2000a), *Causality: Models, Reasoning, and Inference*, Cambridge University Press: Cambridge.

Pearl, Judea (2000b), Comment on "Causal Inference Without Counterfactuals" by A.P. Dawid, *Journal of the American Statistical Association* 95, 428-431.

Pearl, Judea (2001), "Causal Inference in the Health Sciences: A Conceptual Introduction", UCLA Computer Science Dept., Cognitive Systems Laboratory Technical Report R-282.

Pratt, John W. and Robert Schlaiffer (1988), "On the Interpretation and Observation of Laws", *Journal of Econometrics* 39, 23-52.

Quandt, Richard E. (1958), "The Estimation of the Parameters of a Linear Regression System Obeying Two Separate Regimes", *Journal of the American Statistical Association* 53, 873-880.

Quandt, Richard E. (1972), "A New Approach to Estimating Switching Regressions", *Journal of the American Statistical Association* 67, 306-310.

Reichenbach, H. (1956), *The Direction of Time*, Berkeley and Los Angeles: University of California Press.

Robins, James M. (1986), "A New Approach to Causal Inference in Mortality Studies with a Sustained Exposure Period – Application to Control of the Healthy Worker Survivor Effect", *Mathematical Modelling* 7, 1393-1512.

Robins, James M. (1987), "Addendum to ' A New Approach to Causal Inference in Mortality Studies with a Sustained Exposure Period – Application to Control of the Healthy Worker Survivor Effect'", *Comput. Math. Applic.* 14, 923-945.

Robins, James M. (1995), ""Discussion of 'Causal Diagrams for empirical research' by J. Pearl, *Biometrika* 82, 695-698.

Robins, James M. and Sander Greenland (2000), Comment on "Causal Inference Without Counterfactuals" by A.P. Dawid, *Journal of the American Statistical Association* 95, 431-435.

Rosenbaum, Paul R. (1984), "The Consequences of Adjustment for a Concomitant Variable That Has Been Affected by the Treatment", *Journal of the Royal Statistical Society* Series A 147, 656-666.

Rosenbaum, Paul R. (1995a), *Observational Studies*, Springer: New York.

Rosenbaum, Paul R. (1995b), "Discussion of 'Causal Diagrams for empirical research' by J. Pearl, *Biometrika* 82, 698-699.

Rosenbaum, Paul R. and Donald B. Rubin (1983), "The Central Role of the Propensity Score in Observational Studies for Causal Effects", *Biometrika* 70, 41-55.

Rosenbaum, Paul R. and Donald B. Rubin (1984a), "Estimating the Effects Caused by Treatments", Comment on "On the Nature and Discovery of Structure" by J.W. Pratt and R. Schlaifer, *Journal of the American Statistical Association* 79, 26-28.

Rosenbaum, Paul R. and Donald B. Rubin (1984b), "Reducing Bias in Observational Studies using Subclassification on the Propensity Score", *Journal of the American Statistical Association* 79, 516-524.

Roy, A.D. (1951), "Some Thoughts on The Distribution of Earnings", *Oxford Economic Papers* 3, 135-146.

Rubin, Donald B. (1974), "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies", *Journal of Educational Psychology* 66, 688-701.

Rubin, Donald B. (1977), "Assignment to Treatment Group on the Basis of a Covariate", *Journal of Educational Statistics* 2, 1-26.

Rubin, Donald B. (1978), "Bayesian Inference for Causal Effects: The Role of Randomization", Annals of Statistics 6, 34-58.

Rubin, Donald B. (1980), Comment on "Randomization Analysis of Experimental Data: The Fisher Randomization Test" by D. Basu, Journal of the American Statistical Association 75, 591-593.

Rubin, Donald B. (1986), "Which Ifs Have Causal Answers", Comment on Holland (1986), Journal of the American Statistical Association 81, 961-962.

Rubin, Donald B. (1990), "Neyman (1923) and Causal Inference in Experiments and Observational Studies", comment on Neyman (1923), *Statistical Science* 5, 472-480.

Salmon, Wesley (1980), "Probabilistic Causality", *Pacific Philosophical Quarterly* 61, 50-74.

Salmon, Wesley (1998), *Causality and Explanation*, New York and Oxford: Oxford University Press.

Simon, Herbert A. and Nicholas Rescher (1966), "Cause and Counterfactual", *Philosophy of Science* 33, 323-340.

Skyrms, Brian (1984), "EPR: Lessons for Metaphysics", in P. French, T. Uehling, H. Wettstein (eds), *Midwest Studies in Philosophy IX – Causation and Causal Theories*, Minneapolis: University of Minnesota Press.

Skyrms, Brian (1988), "Probability and Causation", *Journal of Econometrics* 39, 53-68.

Sobel, Michael E. (1995), "Causal Inference in the Social and Behavioral Sciences", in G. Arminger, C. C. Clogg, and M. E. Sobel (eds), *Handbook of Statistical Modeling for the Social and Behavioral Sciences*, Plenum Press, New York.

Sosa, Ernest and Michel Tooley (1993a) (eds.), *Causation*, Oxford: Oxford University Press.

Sosa, Ernest and Michel Tooley (1993b), "Introduction", in E. Sosa and M. Tooley (eds), *Causation*, Oxford: Oxford University Press.

Speed, T. J. (1990), "Introductory Remarks on Neyman (1923)", *Statistical Science* 5, 463-464.

Spirtes, P., C. Glymour and R. Scheines (2000), *Causation, Prediction, and Search*, 2nd ed., New York: Springer.

Stalnaker, Robert (1984), *Inquiry*, Boston, MA: Bradford Books.

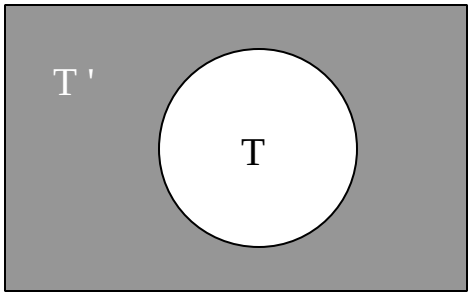
Suppes, Patrick (1970), *A Probabilistic Theory of Causality*, Amsterdam: North Holland.

Suppes, Patrick (1984), "Conflicting Intuitions about Causality", in P. French, T. Uehling, H. Wettstein (eds), *Midwest Studies in Philosophy IX – Causation and Causal Theories*, Minneapolis: University of Minnesota Press.

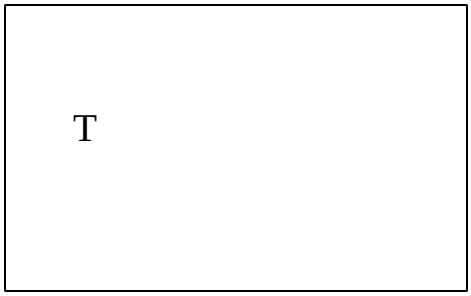
Wright, Sewall (1921), "Correlation and Causation", *Journal of Agricultural Research* 20, 557-585.

Wright, Sewall (1934), "The Method of Path Coefficients", *Annals of Mathematical Statistics* 5, 161-215.

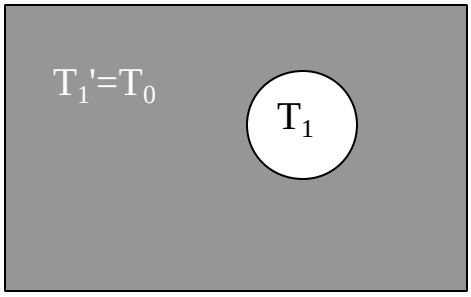
Figure 1. Possible treatment worlds in the POM.



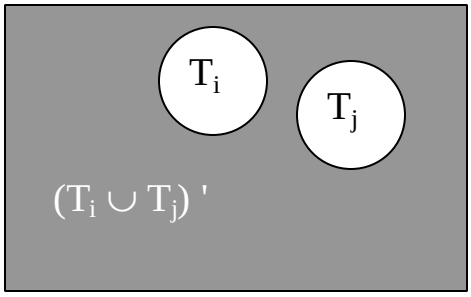
1a.



1b.



1c.



1d.

Table 1. Varieties of Counterfactuals

Number of treatments in T	Treatment of interest	Counterfactual treatment	Causal effect	Interpretation / Notes
M=2	t_1	t_0	$\Delta_{t_1 t_0} = Y_{t_1} - Y_{t_0}$	The null treatment, in most cases the counterfactual of interest. Usually equals t_1', differs only if explicitly specified (as in experimental studies), or if SUTVA is violated.
		t_1'	$\Delta_{t_1 t_1'} = Y_{t_1} - Y_{t_1'}$	*Anything* that is not t_1. Usually applies in observational studies, where it equals t_0.
M>2	t_i	t_0	$\Delta_{t_i t_0} = Y_{t_i} - Y_{t_0}$	The null treatment, again the counterfactual of interest in most cases. Relevant as baseline.
		$t_j \neq t_i$	$\Delta_{t_i t_j} = Y_{t_i} - Y_{t_j} = \Delta_{t_i t_k} - \Delta_{t_j t_k} = \Delta_{t_i t_0} - \Delta_{t_j t_0}$	Any other particular treatment can be used as counterfactual, for interpretation important to note that the baseline (usually the null) is implicit.
		t_i'	$\Delta_{t_i t_i'} = Y_{t_i} - Y_{t_i'} = Y_{t_i} - F(Y_{t_0}, Y_{t_1}, ..., Y_{t_{i-1}}, Y_{t_{i+1}}, ..., Y_{t_{m-1}})$	*Everything* that is not t_i – the *absolute counterfactual*, the outcome of which is given as a function of the outcomes of all treatments except t_i

For M>2, as in the discussion in the text, assume that W=T.

IZA Discussion Papers

No.	Author(s)	Title	Area	Date
280	P. Apps R. Rees	Household Saving and Full Consumption over the Life Cycle	7	04/01
281	G. Saint-Paul	Information Technology and the Knowledge Elites	5	04/01
282	J. Albrecht A. Björklund S. Vroman	Is There a Glass Ceiling in Sweden?	5	04/01
283	M. Hagedorn A. Kaul V. Reinthaler	Welfare Analysis in a Schumpeterian Growth Model with Capital	7	04/01
284	H. Rapoport A. Weiss	The Optimal Size for a Minority	1	04/01
285	J. Jerger C. Pohnke A. Spermann	Gut betreut in den Arbeitsmarkt? Eine mikroökonometrische Evaluation der Mannheimer Arbeitsvermittlungsagentur	5	04/01
286	M. Fertig C. M. Schmidt	First- and Second-Generation Migrants in Germany – What Do We Know and What Do People Think	1	04/01
287	P. Guggenberger A. Kaul M. Kolmar	Efficiency Properties of Labor Taxation in a Spatial Model of Restricted Labor Mobility	3	04/01
288	D. A. Cobb-Clark	Getting Ahead: The Determinants of and Payoffs to Internal Promotion for Young U.S. Men and Women	5	04/01
289	L. Cameron D. A. Cobb-Clark	Old-Age Support in Developing Countries: Labor Supply, Intergenerational Transfers and Living Arrangements	3	04/01
290	D. A. Cobb-Clark M. D. Connolly C. Worswick	The Job Search and Education Investments of Immigrant Families	1	04/01
291	R. T. Riphahn	Cohort Effects in the Educational Attainment of Second Generation Immigrants in Germany: An Analysis of Census Data	1	05/01

292	E. Wasmer	Between-group Competition in the Labor Market and the Rising Returns to Skill: US and France 1964-2000	5	05/01
293	D. Cobb-Clark T. F. Crossley	Gender, Comparative Advantage and Labor Market Activity in Immigrant Families	1	05/01
294	Š. Jurajda	Estimating the Effect of Unemployment Insurance Compensation on the Labor Market Histories of Displaced Workers	3	05/01
295	F. Duffy P. P. Walsh	Individual Pay and Outside Options: Evidence from the Polish Labour Force Survey	4	05/01
296	H. S. Nielsen M. Rosholm N. Smith L. Husted	Intergenerational Transmissions and the School-to-Work Transition of 2nd Generation Immigrants	1	05/01
297	J. C. van Ours J. Veenman	The Educational Attainment of Second Generation Immigrants in The Netherlands	1	05/01
298	P. Telhado Pereira P. Silva Martins	Returns to Education and Wage Equations	5	06/01
299	G. Brunello C. Lucifora R. Winter-Ebmer	The Wage Expectations of European College Students	5	06/01
300	A. Stutzer R. Lalive	The Role of Social Work Norms in Job Searching and Subjective Well-Being	5	06/01
301	J. R. Frick G. G. Wagner	Economic and Social Perspectives of Immigrant Children in Germany	1	06/01
302	G. S. Epstein A. Weiss	A Theory of Immigration Amnesties	1	06/01
303	G. A. Pfann B. F. Blumberg	Social Capital and the Uncertainty Reduction of Self-Employment	5	06/01
304	P. Cahuc E. Wasmer	Labour Market Efficiency, Wages and Employ-ment when Search Frictions Interact with Intra-firm Bargaining	2	06/01
305	H. Bonin	Fiskalische Effekte der Zuwanderung nach Deutschland: Eine Generationenbilanz	1	06/01

306	H. Bonin G. Abío E. Berenguer J. Gil C. Patxot	Is the Deficit under Control? A Generational Accounting Perspective on Fiscal Policy and Labour Market Trends in Spain	2	06/01
307	G. A. Pfann	Downsizing	1/5	06/01
308	G. A. Pfann D. S. Hamermesh	Two-Sided Learning, Labor Turnover and Worker Displacement	1	06/01
309	G. Brunello	On the Complementarity between Education and Training in Europe	5	06/01
310	U. Sunde	Human Capital Accumulation, Education and Earnings Inequality	5	06/01
311	G. Brunello	Unemployment, Education and Earnings Growth	3	06/01
312	C. Furnée M. Kemler G. A. Pfann	The Value of Pain Relief	5	06/01
313	A. Ferrer-i-Carbonell B. M.S. van Praag	The Subjective Costs of Health Losses due to Chronic Diseases: An Alternative Model for Monetary Appraisal	7	06/01
314	B. M.S. van Praag A. Ferrer-i-Carbonell	Age-Differentiated QALY Losses	7	06/01
315	W. H. J. Hassink R. Schettkat	On Price-Setting for Identical Products in Markets without Formal Trade Barriers	7	06/01
316	M. Frondel C. M. Schmidt	Rejecting Capital-Skill Complementarity at all Costs	5	06/01
317	R. Winkelmann	Health Care Reform and the Number of Doctor Visits – An Econometric Analysis	7	06/01
318	M. Pannenberg G. G. Wagner	Overtime Work, Overtime Compensation and the Distribution of Economic Well-Being: Evidence for West Germany and Great Britain	1	06/01
319	R. Euwals R. Winkelmann	Why do Firms Train? Empirical Evidence on the First Labour Market Outcomes of Graduated Apprentices	1	06/01

320	R. Fahr U. Sunde	Strategic Hiring Behavior in Empirical Matching Functions	1	06/01
321	P. Telhado Pereira P. Silva Martins	Is there a Return – Risk Link in Education?	5	07/01
322	O. Hübler U. Jirjahn	Works Councils and Collective Bargaining in Germany: The Impact on Productivity and Wages	1	07/01
323	A. Frederiksen E. K. Graversen N. Smith	Overtime Work, Dual Job Holding and Taxation	1	07/01
324	M. Pflüger	Trade, Technology and Labour Markets: Empirical Controversies in the Light of the Jones Model	2	07/01
325	R. A. Hart J. R. Malley U. Woitek	Real Wages and the Cycle: The View from the Frequency Domain	1	07/01
326	J. S. Earle Á. Telegdy	Privatization and Productivity in Romanian Industry: Evidence from a Comprehensive Enterprise Panel	4	07/01
327	H. Gersbach A. Schmutzler	A Product Market Theory of Training and Turnover in Firms	5	07/01
328	F. Breyer	Why Funding is not a Solution to the “Social Security Crisis”	3	07/01
329	X. Gong A. van Soest	Wage Differentials and Mobility in the Urban Labor Market: A Panel Data Analysis for Mexico	1	07/01
330	D. N. Margolis K. G. Salvanes	Do Firms Really Share Rents with Their Workers?	5	07/01
331	R. Winkelmann	Why Do Firms Recruit Internationally? Results from the IZA International Employer Survey 2000	5	07/01
332	M. Rosholm	An Analysis of the Processes of Labour Market Exclusion and (Re-) Inclusion	3	07/01
333	W. Arulampalam R. A. Naylor J. P. Smith	A Hazard Model of the Probability of Medical School Dropout in the United Kingdom	5	07/01

334	P. A. Puhani	Wage Rigidities in Western Germany? Microeconometric Evidence from the 1990s	1	07/01
335	R. Fahr U. Sunde	Disaggregate Matching Functions	1	07/01
336	F. Lima P. Telhado Pereira	Careers and Wage Growth within Large Firms	5	07/01
337	F. Büchel M. Pollmann-Schult	Overeducation and Skill Endowments: The Role of School Achievement and Vocational Training Quality	5	08/01
338	C. Bell H. Gersbach	Child Labor and the Education of a Society	5	08/01
339	A. Ibourk B. Maillard S. Perelman H. R. Sneessens	The Matching Efficiency of Regional Labour Markets: A Stochastic Production Frontier Estimation, France 1990-1995	1	08/01
340	X. Wauthy Y. Zenou	How Does Imperfect Competition in the Labor Market Affect Unemployment Policies?	3	08/01
341	S. Kohns	Testing for Asymmetry in British, German and US Unemployment Data	1	08/01
342	W. Schnedler	The Virtue of Being Underestimated: A Note on Discriminatory Contracts in Hidden Information Models	5	08/01
343	H. Bonin	Will it Last? An Assessment of the 2001 German Pension Reform	3	08/01
344	E. Plug P. Berkhout	Effects of Sexual Preferences on Earnings in the Netherlands	5	08/01
345	J. Hampe M. Steininger	Survival, Growth, and Interfirm Collaboration of Start-Up Companies in High-Technology Industries: A Case Study of Upper Bavaria	5	08/01
346	L. Locher	The Determination of a Migration Wave Using Ethnicity and Community Ties	1	08/01
347	M. Lofstrom F. D. Bean	Labor Market Conditions and Post-Reform Declines in Welfare Receipt Among Immigrants	3	08/01
348	S. Neuman A. Ziderman	Can Vocational Education Improve the Wages of Minorities and Disadvantaged Groups? The Case of Israel	5	08/01

349	J. T. Addison P. Portugal	Job Search Methods and Outcomes	1	08/01
350	J. T. Addison P. Portugal	Unemployment Duration: Competing and Defective Risks	1	08/01
351	J. D. Brown J. S. Earle	Gross Job Flows in Russian Industry Before and After Reforms: Has Destruction Become More Creative?	4	08/01
352	J. T. Addison J. S. Heywood X. Wei	Unions and Plant Closings in Britain: New Evidence from the 1990/98 WERS	1	08/01
353	T. Bauer S. Bender	Flexible Work Systems and the Structure of Wages: Evidence from Matched Employer-Employee Data	5	08/01
354	J. Kluve	On the Role of Counterfactuals in Inferring Causal Effects of Treatments	6	09/01