Cover Letter: A cover letter must accompany the submitted manuscript stating that the paper and the data have not previously been published, either in whole or in part (unless as an abstract), and that no similar paper is in press or under review elsewhere.

The cover letter should also state potential conflicts of interest that could raise questions about a paper's credibility if disclosed later. (See above.)

Any closely-related papers by the authors (including manuscripts that are published, in press, or under review) should be sent to the journal editorial office, as described above. The cover letter should make clear the independent contribution of the submitted paper.

On the Consistency Rule in Causal Inference:

An Axiom, Definition, Assumption, or a Theorem?*

Judea Pearl

University of California, Los Angeles

Computer Science Department

Los Angeles, CA, 90095-1596, USA

judea@cs.ucla.edu

type of manuscript: pick one

appreciation, brief report, book and software review, commentary, editorial, editorial post script, epidemiology & society, errata, filter hypothesis, letter, original article, poetry and epidemiology, rememberance, review article, vignette, virtual epidemiology, voices

*I am grateful to Sander Greenland for bringing this topic to my attention and for his advice in writing this note. This research was partially supported by grants from NIH #1R01 LM009961-01, NSF #IIS-0914211, and ONR #N000-14-09-1-0665.

fyi - refs need to numbered in order of appearance and superscript in text. i put only numbers were ok, but left names/years where it wouldn't make sense if removed. let me know if i should remove the years in the text.

Abstract

In two recent communications, Cole and Frangakis (2009)[1] and VanderWeele (2009)[2] conclude that the consistency rule used in causal inference is an assumption that precludes any side-effects of treatment/exposure modalities on the outcomes of interest. They further develop auxiliary notation to make this assumption formal and explicit. I argue that the consistency rule is a theorem in the logic of counterfactuals and need not be altered. Instead, warnings of potential side-effects should be embodied in modeling practices that facilitate explicit and transparent encoding of all causal assumptions.

1 Introduction

Informally, the consistency rule states that an individual's potential outcome under a hypothetical condition that happened to materialize is precisely the outcome experienced by that individual. When expressed formally, this rule reads:[3]

$$X(u) = x \implies Y_x(u) = Y(u) \tag{1}$$

and serves as one of most basic mathematical tools for deriving causal effects from statistical data. Since any mathematical derivation must rest on a formal system of axioms, models, interpretations and inference rules, the status of the consistency rule can best be elucidated by examining its role in formal theories of actions and counterfactuals.

2 The Possible Worlds Account

Robert Stanalker (1972)[4] and David Lewis (1973)[5], the philosophers who first developed such formal theories, gave a "possible worlds" interpretation to action and counterfactual

sentences. In their account, the action sentence "If we paint the wall red my uncle will be cheerful," is equivalent to an "as if" counterfactual sentence: "if the wall were red, my uncle would be cheerful." Such sentence is deemed true if the "closest world" satisfying the antecedent proposition "the wall is red" also satisfies the consequent proposition: "my uncle is cheerful." The "similarity" measure that ranks worlds for closeness can be quite general, and requires only that every world be closest to itself.

If an analyst believes that different ways of performing action A are likely to have different effects on the outcome(s), the analyst must specify the conditions that characterize each nuance, and what differences they make to other variables in the model. For example, if a certain type of paint tends to produce toxic vapor, a specific nuance of the action "paint the wall red" would read: "the wall is red and there is toxic vapor in my uncle's room" while another would read: "the wall is red and there is no toxic vapor in my uncle's room" The antecedent A of the counterfactual sentence "if A were true then B" would then be conjunctions of the primary effect of the action (red wall) and its secondary effects (toxic vapor). Naturally, the model must further explicate how each conjunction affects the outcome of interest, e.g., "my uncle being cheerful." These are encoded through the "similarity" measure that renders some worlds more similar than others and thus determines the likely outcomes of each action.

Lewis's (1973)[5] "closest world" interpretation of counterfactuals entails certain universal properties, called theorems, that hold true regardless of the similarity measure used in ranking worlds. One such theorem is the consistency rule, first stated formally in Gibbard and Harper (1976, p. 156).[6] It reads as follows: For all A and B, if A is true, then if B would have prevailed (counterfactually) had A been true, it must be true already. This

may sound tautological, but when translated into experimental setting, it usually evokes reservations, for it reads: "a person who chose treatment $X = x$ and recovered would also have recovered in a clinical trial if assigned treatment x by design." Here we become immediately suspicious of possible side-effects that the experimental protocol might have on recovery which, if significant, would seems to invalidate the consistency rule. Not so. According to Lewis's theory, the existence of such side-effects should merely modify the proposition "*treatment* $= x$" to include the additional conditions imposed by the treatment (e.g., toxic vapors in the case of wall painting, psychological stress in the case of clinical trials) to ensure that the counterfactual antecedent A represents the relevant features of the treatment actually received.

3 The Structural Account

While Lewis's "closest world" account may seem esoteric to practicing researchers, the structural account of counterfactuals (Pearl, 2000, Chapter 7)[7] should make this argument more transparent. The latter is based not on metaphysical notions of "similarity" and "possible worlds," but on the physical mechanisms that govern our world. In this account, a "model" M is encoded as a collection of functions, each representing a physical mechanism responsible for assigning a value to a distinct variable in the model. The value assigned depends on values previously taken by other variables in the model and on a vector U of features that characterize each experimental unit u. The definition of counterfactuals $Y_x(u)$ in this model is based on solving the equations in a modified version of M, called M_x, and

it reads:

$$Y_x(u) \triangleq Y_{M_x}(u). \tag{2}$$

In words, the value that outcome Y would take in unit u, had X been x is given by the solution for Y in a "modified" model M_x in which the equation for X is replaced by the equation $X = x$. The modified model M_x represents the "least invasive" perturbation of M necessary for enforcing the condition $X = x$ prescribed by the antecedent of the counterfactual.

Having a model M and a formal definition for counterfactuals (2) enables us to assign a truth value to any statement involving counterfactuals or joint probabilities of counterfactuals. In particular, this definition also enables us to derive *theorems*, namely, counterfactual statements that hold true in *all* models M, regardless of the content of the equations or their organization. Not surprisingly, the consistency rule articulated in (1) can be shown to be among those theorems.[8,9]

This agreement between two diverse accounts of counterfactuals is not coincidental; the structural account can be given a "closest world" interpretation, provided worlds that share identical histories are deemed equally similar.[7]

4 Discussion

The implications of the last two sections are as follows: Contrary to the conclusion of Cole, Frangakis, and VanderWeele, the logic of counterfactuals tolerates no departure from the consistency rule and, therefore, there is no assumption conveyed by the rule. Considerations of side-effects are embodied in the standard modeling requirement that the action-defining

proposition, $X = x$, properly describes the conditions created by a given treatment (or exposure).

When models are transparent, this translates into an even milder requirement that a model should make no claim which the analyst finds objectionable. Figure 1 depicts two models for the action statement: "If we paint the wall red my uncle will be cheerful." Figure 1(a) disregards the possibility that some paints may release toxic vapor, and Fig. 1(b) explicitly displays this possibility. Readers versed in causal diagrams (Greenland et al., 1999)[10] will recognize immediately that, if the analyst deems toxic paint to be a likely outcome of the action, Fig. 1(a) is not merely incomplete, but makes blatantly false claims. It claims, for example, that my uncle's mood is independent of the action, given wall color. Assumptions, in language of diagrams, are encoded not in the arrows, but in the missing arrows, hence the arrow missing between "action" and "mood" vividly displays a false premise, one that is rectified in Fig. 1(b).

A natural question to ask is how the consistency rule is positioned in the "potential outcome" framework of Neyman (1923)[11] and Rubin (1974)[12];[1] is it a definition, an axiom, an assumption or a theorem?

footnotes are not allowed - add back to text?

Unlike the "possible worlds" and structural accounts, the potential outcome framework does not *define* counterfactuals but takes them as primitive, undefined quantities. It is the

[1] By "potential outcome" framework I mean the counterfactual theory introduced by Neyman (1923),[11] Wilks (1955),[13] and Rubin (1974),[12] which considers causal inference to be a statistical "missing value" problem and makes no reference to possible worlds, structural equations or causal diagrams. One should note that the subscript notation in Eq. (1) and the notion of "potential outcome" is shared by many analysts who do rely on structural equations and diagrams (e.g., Glymour and Greenland (2008)[14] and VanderWeele and Robins (2007)[15]).

consistency rule alone, often written

$$Y = xY_1 + (1 - x)Y_0 \tag{3}$$

that connects the undefined primitives, Y_0 and Y_1, to observed quantities, X and Y, and gives the former empirical meaning. Absent this rule, the variables Y_1 and Y_0 would play no role in causal effects estimation or in any decision making application.

Thus, while the structural and possible worlds accounts derive the consistency rule from formal definitions of counterfactuals, the potential outcome framework reverses the logic and uses the consistency rule to define counterfactuals. In this role, the consistency rule acts as a self-evident axiom, rather than a theorem or an assumption. How self evident it is, depends on the context and application. As noted in Cole and Frangakis, the consistency rule can be taken as self-evident in an ideal experimental study, where investigators exercise full control over the treatment conditions. Whether it retains its self-evident status in observational studies becomes a matter of faith, or an *assumption* which, given the undefined status of counterfactuals in this framework, can benefit indeed from the explication offered by Cole, Frangakis, and VanderWeele.

In the formal frameworks of possible worlds and structural models, however, these assumptions are explicated in a different form and in a different phase of the analysis. The task of ensuring that all relevant side-effects are accounted for is solely the responsibility of the modeller and, assuming the modeler upholds this responsibility, the analyst can safely use the simple, unmodified version of the rule, as in Eq. (2). The need to alert investigators to possible side effects of treatment modalities, an important requirement in the potential outcome framework, also underscores the importance of separating modeling assumption

from definitions and rules of inference.

References

[1] S.R. Cole and C.E. Frangakis. The consistency statement in causal inference: A definition or an assumption? *Epidemiology*, 20:3–5, 2009.

[2] T.J. VanderWeele. Concerning the consistency assumption in causal inference. *Epidemiology*, 20(1):880–883, 2009.

[3] J.M. Robins. A new approach to causal inference in mortality studies with a sustained exposure period – applications to control of the healthy workers survivor effect. *Mathematical Modeling*, 7:1393–1512, 1986.

[4] R.C. Stalnaker. Letter to David Lewis, 1972. In W.L. Harper, R. Stalnaker, and G. Pearce (Eds.), *Ifs*, D. Reidel, Dordrecht, pages 151–152, 1981.

[5] D. Lewis. Counterfactuals and comparative possibility, 1973. In W.L. Harper, R. Stalnaker, and G. Pearce (Eds.). *Ifs*, D. Reidel, Dordrecht, pages 57–85, 1981.

[6] A. Gibbard and L. Harper. Counterfactuals and two kinds of expected utility, 1976. In W.L. Harper, R. Stalnaker, and G. Pearce (Eds.), *Ifs*, D. Reidel, Dordrecht, pages 153–169, 1981.

[7] J. Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, New York, 2000. Second ed., 2009.

[8] D. Galles and J. Pearl. An axiomatic characterization of causal counterfactuals. *Foundation of Science*, 3(1):151–182, 1998.

[9] J.Y. Halpern. Axiomatizing causal reasoning. In G.F. Cooper and S. Moral, editors, *Uncertainty in Artificial Intelligence*, pages 202–210. Morgan Kaufmann, San Francisco, CA, 1998. Also, *Journal of Artificial Intelligence Research* 12:3, 17–37, 2000.

[10] S. Greenland, J. Pearl, and J.M Robins. Causal diagrams for epidemiologic research. *Epidemiology*, 10(1):37–48, 1999.

[11] J. Neyman. On the application of probability theory to agricultural experiments. Essay on principles. Section 9. *Statistical Science*, 5(4):465–480, 1923.

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

[13] M.B. Wilks. The randomization analysis of a generalized randomized block design. *Biometrik*, 42(1/2):70–79, 1955.

[14] M.M. Glymour and S. Greenland. Causal diagrams. In K.J. Rothman, S. Greenland, and T.L. Lash, editors, *Modern Epidemiology*, pages 183–209. Lippincott Williams & Wilkins, Philadelphia, PA, 3rd edition, 2008.

[15] T.J. VanderWeele and J.M. Robins. Four types of effect modification: A classification based on directed acyclic graphs. *Epidemiology*, 18(5):561–568, 2007.

fig on separate page

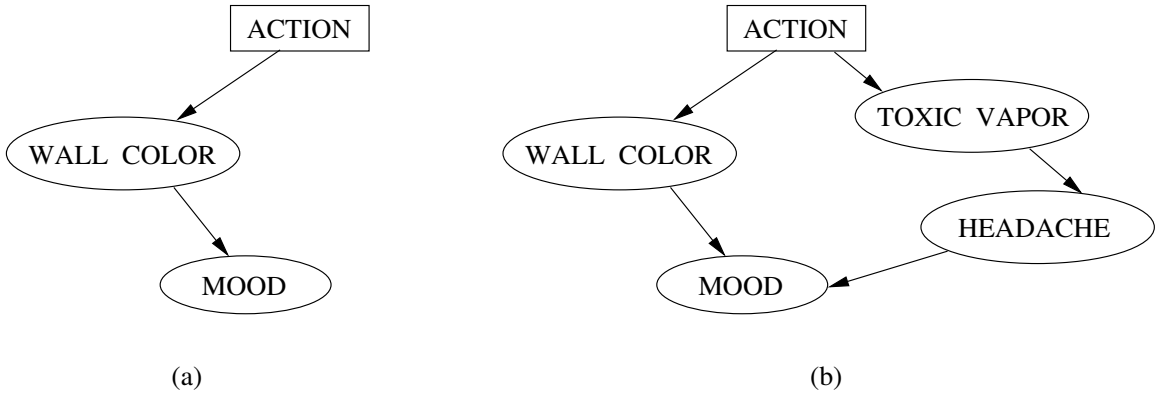

Figure 1: Two models interpreting the action phrase "paint the wall red." (a) Neglects the side effect "toxic vapor," shown in (b).