In *Epidemiology*, Vol. 21(6):872-875, 2010.

TECHNICAL REPORT
R-358
August 2010

On the Consistency Rule in Causal Inference: Axiom, Definition, Assumption, or Theorem?

Judea Pearl
University of California, Los Angeles
Computer Science Department
Los Angeles, CA, 90095-1596, USA
Telephone: 310.825.3243
Fax: 310.794.5057
judea@cs.ucla.edu

Abstract

Remove the abstract (and move to the Main Text). In two recent communications, Cole and Frangakis and VanderWeele conclude that the consistency rule used in causal inference is an assumption that precludes any side-effects of treatment/exposure 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, even in cases where different versions of treatment/exposure have side effects on the outcome. Instead, warnings of potential side-effects should be embodied in standard modeling practices, using graphs or structural equation models, in which causal assumptions are given transparent and unambiguous representation.

In two recent communications, Cole and Frangakis[1] and VanderWeele[2] interpret the consistency rule of causal inference an assumption that is violated whenever versions of treatment/exposure have unintended side effects on the outcomes of interest. They further develop auxiliary notation to make this assumption formal and explicit. I argue that the consistency rule is in fact a theorem in the logic of counterfactuals and need not be altered. Side-effects can be made explicit in standard modeling practices, whenever transparent models are used.

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

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

where $X(u)$ stands for the exposure that person u actually experienced, $Y_x(u)$ the potential outcome of person u had the exposure been at level $X = x$, and $Y(u)$ is the outcome actually realized by u. As a mathematical tool, the consistency rule permits us to write (for any z):

$$P(Y_x = y|Z = z, X = x) = P(Y = y|Z = z, X = x),$$

thus converting expressions involving probabilities of counterfactuals to expressions involving ordinary conditional probabilities of measured variables. Most theoretical results in causal inference, including those invoking "ignorability" assumptions, the control of confounding and the validity of propensity scores methods, owe their derivations to the consistency rule.

Because 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.

The Possible-Worlds Account

Robert Stalnaker[4] and David Lewis,[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. In our example, every world entailing "toxic vapors" will also entail "my uncle is far from cheerful," and will be placed closer to ours than any world in which "my uncle feels cheerful."

Lewis's[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 by Gibbard and Harper.[6,p. 156] 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 then 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, side-effects that, if significant, would seem 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.

The Structural Account

While Lewis's "closest-world" account may seem esoteric to practicing researchers, the structural account of counterfactuals[7,*ch.* 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, as perceived by the modeller. In this account, a "model" M embodies 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.

In practice, it is extremely rare that one would be able to specify the functional relationships among the variables, or even list the variables involved; partial, qualitative and provisional knowledge of these relationships is all that one can hope to encode in a model. There is also no guarantee that the scanty knowledge encoded in the model is free of errors, as, for example, when we neglect to encode the possibility of "toxic vapor" in the wall-painting example. However, having a formal model ensures that we make consistent and maximum use of the knowledge that we do select to encode in the model.

In particular, having a model M and a formal definition for counterfactuals (2) enables us to assign a truth value to any statement involving counterfactuals, as well as joint probabilities of counterfactuals. Such assignments enable us to determine if the knowledge encoded in a partially-specified model is sufficient for drawing specific types of causal conclusions from the data.[8] More importantly, 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.[9,10]

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]

Discussion

The implications of the last two sections are that 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 that the analyst finds objectionable. The Figure depicts two models for the action statement: "If we paint the wall red my uncle will be cheerful." Figure A disregards the possibility that some paints may release toxic vapor, and Figure B explicitly displays this possibility. Readers versed in causal diagrams[11,12] will recognize immediately that, if the analyst deems toxic paint to be a likely outcome of the action, Figure 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 the 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 Figure B.

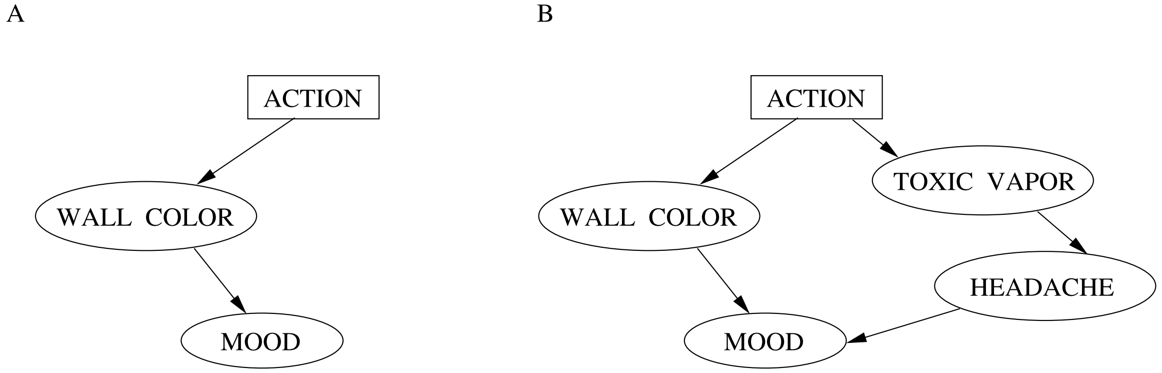

Figure: Two models interpreting the action phrase "paint the wall red." (A) Neglects the side effect "toxic vapor," which is shown in (B).

A natural question to ask is how the consistency rule is positioned in the "potential-outcome" framework of Neyman,[13] Wilks,[14] and Rubin[15] – in which causal inference is considered to be a statistical "missing value" problem, bearing no relation to possible worlds, structural equations or causal diagrams. Is it a definition, an axiom, an assumption or a theorem?

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 consistency rule alone, often written as

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

that connects the undefined primitives, Y_0 and Y_1, to observed quantities, X and Y, and endows the former with empirical meaning. Absent this rule, the variables Y_1 and Y_0 would bear no relation to any measured quantity save for the verbal, informal understanding that each stands for the "potential outcome" of a unit under some unspecified conditions indexed by the subscripts 1 and 0.

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 as a theorem or an assumption. How self-evident it is depends on the context and application. As noted by Cole and Frangakis,[1] the consistency rule appears to be compelling in ideal experiments where investigators are presumed to have full control, and full awareness, of all treatment conditions. Practical experimental designs, however, cannot guarantee such control, and the need invariably arises to enumerate the conditions indexed by subscripts 1 and 0 in Equation (3). This occurs whenever we venture to transport conclusions of one study to a new experimental setup characterized by somewhat different conditions (e.g., a wider population), and to argue that the differences are irrelevant. Whether the consistency rule retains its self-evident status in this transport becomes a matter of faith, or an assumption, which may benefit from the explication offered by Cole and Frangakis[1] and byVanderWeele.[2]

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 practitioner-modeller and, assuming the modeller upholds this responsibility, the analyst can safely use the simple, unmodified version of the rule, as in Eq. (2). Separating modeling assumptions from definitions and rules of inference has the advantage of freeing the rules from the subtleties of the assumptions.

Conclusions

I agree with Cole and Frangakis[1] and with VanderWeele[2] that assumptions of "no side-effects" need be attended to with utmost diligence, that they deserve a formal representation, and that no representation, however sophisticated, can capture side-effect assumptions that researchers fail to notice or acknowledge.

I also agree that it is practically not possible to account analytically for all the different ways in which an exposure of level x can be given. I argue, however, that, if one possesses experience about what ways of giving exposure x can be considered similar and what ways cannot, such experience should be encoded not by altering the consistency rule but, rather, in the same model in which other causal assumptions are encoded. This model can take the form of a causal diagram, in which assumptions receive vivid and unambiguous representation or, if one prefers algebraic notation, through counterfactual formulae of the "ignorability" type. The latter two representations are logically equivalent,[8,16(pp. 98−−102)] and differ only in emphasis and transparency.

I further argue that the distinction between an "assumption" and a "theorem" is not just a matter of semantics, but rather carries profound implications in research, communication and education, not unlike the implications of labeling the Pythagorean Theorem as a "theorem," not an "assumption." Although right-angle triangles hardly exist in the practical world, the label "theorem" serves useful purposes to geometers, astronomers, and engineers. First, it gives mathematicians the license to communicate results using a few standard, albeit ideal, mathematical objects, (e.g., straight lines, right-angles) rather than the much larger space of deviants from the ideal. Second, it gives mathematicians the freedom to explore properties of more intricate objects (e.g., polygons, spherical geometry, calculus) while delegating the task of assessing the practical applicability of such properties to those who are more intimately familiar with the details of each specific application. Finally, a "theorem" conveys to practitioners the comfortable presence of a solid coherent science behind their practice and the assurance that this science can be relied upon for guidance despite its dealing with ideal mathematical objects.

The science of counterfactuals, like that of geometry, deals with ideal mathematical objects such as local interventions, indexed by a finite set $X = x$ of propositions, and counterfactuals defined by such local interventions in accordance with Eq. (2). Practicing epidemiologists would do well to acquire the tools developed by the science of counterfactuals, despite the ideal nature of its premises. The label "theorem" acknowledges the consistency of that science; the label "assumption" denies its existence.

Acknowledgments

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

References

[1] Cole, SR, Frangakis, CE. The consistency statement in causal inference: A definition or an assumption? *Epidemiology*, 20:3–5, 2009.

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

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

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

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

[8] Pearl, J. An introduction to causal inference. *The International Journal of Biostatistics*, 6(2):DOI: 10.2202/1557–4679.1203, <http://www.bepress.com/ijb/vol6/iss2/7/>, 2010.

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

[10] Halpern, JY. 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.

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

[12] Glymour, MM, Greenland, S. 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.

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

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

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

[16] Pearl, J. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, New York, second edition, 2009.