Concerning the Consistency Assumption in Causal Inference

Tyler J. VanderWeele

Abstract: Cole and Frangakis (*Epidemiology*. 2009;20:3–5) introduced notation for the consistency assumption in causal inference. I extend this notation and propose a refinement of the consistency assumption that makes clear that the consistency statement, as ordinarily given, is in fact an assumption and not an axiom or definition. The refinement is also useful in showing that additional assumptions (referred to here as treatment-variation irrelevance assumptions), stronger than those given by Cole and Frangakis, are in fact necessary in articulating the ordinary assumptions of ignorability or exchangeability. The refinement furthermore sheds light on the distinction between intervention and choice in reasoning about causality. A distinction between the range of treatment variations for which potential outcomes can be defined and the range for which treatment comparisons are made is discussed in relation to issues of nonadherence. The use of stochastic counterfactuals can help relax what is effectively being presupposed by the treatment-variation irrelevance assumption and the consistency assumption.

Cole and Frangakis[1] have provided an interesting commentary on the consistency assumption in the causal inference literature. I would like to raise a number of additional points.

REFINEMENTS IN THE ARTICULATION OF AND NOTATION FOR THE CONSISTENCY ASSUMPTION

Cole and Frangakis define $Y_j(x, k)$ as the potential outcome for individual j if exposure X is set to value x by means k. They let X_j and Y_j^{obs} denote, respectively, the observed values of X and Y for individual j. The consistency assumption is then reformulated as the assumption that "$Y_j^{obs} = Y_j(x, k)$ if $x = X_j$, no matter the value of k." One issue that is not addressed in their commentary is that the range of k will generally vary with x and we will in general need to index treatment variations by k_x, which may be different for each treatment option x. For example, in comparing surgery ($x = 1$) and chemotherapy ($x = 0$) as cancer treatments, the range of treatment variations for surgery (eg, which specific surgeon) differs considerably from the range of treatment variations for chemotherapy (eg, dose). Similarly, one might imagine countless variations on the implementation of a job training program for participants, but there are arguably far fewer variations for a control arm of "not participating." Some aspects of treatment variation (eg, specific hospital, time of treatment initiation) may be common to the treatments being compared but generally not all will be. In addition to specifying the components of k_x, one might also wish to specify some set of values K_x and require $Y_j^{obs} = Y_j(x, k_x)$ when $x = X_j$ only for $k_x \in K_x$, rather than necessarily requiring this consistency condition for all k_x.

With this additional notation, I would like to propose a further refinement. Consider the following 2 conditions. First, whether for each x,

$$Y_j(x, k_x) = Y_j(x, k_x') \text{ for all } k_x, k_x' \in K_x. \text{ (C1)}$$

If condition ($C1$) holds, then, for any $k_x \in K_x$, we could define $Y_j(x)$ as $Y_j(x) := Y_j(x, k_x)$. Second, we might be interested in whether for each j,

$$\text{for some } k_x \in \mathrm{K_x},\ \mathrm{Y}_j^{obs} = Y_j(x, k_x) \text{ when } x = X_j. \text{ (C2)}$$

The condition ($C1$) requires for each x that the potential outcomes $Y_j(x, k_x)$ take the same value irrespective of what means k_x is used to set X to x so long as $k_x \in K_x$. Condition ($C2$) then requires that for some k_x the potential outcome $Y_j(x, k_x)$ is equal to the observed outcome Y_j^{obs} when $x = X_j$. If ($C1$) and ($C2$) hold then for all $k_x \in K_x$, the potential outcome $Y_j(x, k_x)$ is equal to the observed outcome Y_j^{obs} when $x = X_j$.

Several consequences follow from this refinement of the consistency assumption. First, we can better understand how the consistency statement, as ordinarily articulated, is an assumption, and not an axiom or definition. Ordinarily the consistency assumption is articulated as $Y_j^{obs} = Y_j(x)$ when $x = X_j$. Now if condition ($C1$) is satisfied, we can define $Y_j(x) := Y_j(x, k_x)$ for any k_x. Condition ($C2$) then simply becomes $Y_j^{obs} = Y_j(x)$ when $x = X_j$, the ordinary consistency assumption. I would argue that when researchers in the past made the consistency assumption, 2 things were being assumed. First, it was assumed that the notation $Y_j(x)$ was justified ie, that some range of potential interventions under

Submitted 22 December 2008; accepted 14 April 2009.
From the Departments of Epidemiology and Biostatistics, Harvard School of Public Health, Boston, MA.
Correspondence: Tyler J. VanderWeele, Departments of Epidemiology and Biostatistics, Harvard School of Public Health, 677 Huntington Ave, Boston, MA 02115. E-mail: tvanderw@hsph.harvard.edu.

ISSN: 1044-3983/09/2006-0880
DOI: 10.1097/EDE.0b013e3181bd5638

consideration to set X to x would all yield the same potential outcomes. This is what is articulated above as assumption ($C1$); the notation itself presupposed ($C1$). Second, the consistency assumption, stated as $Y_j^{obs} = Y_j(x)$ when $x = X_j$, was assumed and required ($C2$).

The very use of the notation $Y_j(x)$ conceals assumption ($C1$). A similar issue arises in discussions of interference. The notation $Y_j(x)$ is justified only if the outcome for one individual j does not depend on the treatments received by other individuals. This is sometimes referred to as the "no-interference" assumption. Several authors[2–6] have relaxed this assumption by introducing notation such as $Y_j(x_j, \mathbf{x}_{-j})$ to denote the potential outcome for individual j if individual j received treatment x_j and if the individuals other than j had received treatments corresponding to the vector $\mathbf{x}_{-j}$. The no-interference assumption can then be articulated as $Y_j(x_j, \mathbf{x}_{-j}) = Y_j(x_j, \mathbf{x'}_{-j})$ for all $\mathbf{x}_{-j}, \mathbf{x'}_{-j}$. Most work in causal inference does not formally introduce this notation to make the no-interference assumption; the assumption is simply implicit in the ordinary use of the notation $Y_j(x)$. Notation can conceal substantive assumptions. In the case of the consistency assumption, I would argue that the notation $Y_j(x)$ presupposes assumption ($C1$); the consistency assumption itself, stated as $Y_j^{obs} = Y_j(x)$ when $x = X_j$, is then the assumption ($C2$).

A second consequence of this refinement of the consistency assumption concerns the values of x for which ($C1$) is to hold. Cole and Frangakis require that "$Y_j^{obs} = Y_j(x, k)$ if $x = X_j$, no matter the value of k." However, once we follow Cole and Frangakis and introduce the notation $Y_j(x, k_x)$ for the potential outcome for individual j if exposure X is set to value x by means k_x, the question arises whether ($C1$) holds for $x \neq X_j$. The drawing of causal inferences often makes use not only of the consistency assumption but also, as noted by Cole and Frangakis, of the "exchangeability" or "ignorability" assumption. This assumption is often articulated as the independence of the potential outcome $Y_j(x)$ and actual treatment X_j, conditional on some set of confounding variables Z. The very statement of this ignorability assumption requires condition ($C1$) but does not require condition ($C2$). Similarly, assumption ($C1$), but not ($C2$), would be required for causal effects of the form $Y_j(1) - Y_j(0)$ to be well-defined. Importantly, to state the ignorability assumption or to define causal effects such as $Y_j(1) - Y_j(0)$, assumption ($C1$) must hold for values of x other than the observed value X_j. The consistency assumption as articulated by Cole and Frangakis does not imply that ($C1$) holds for $x \neq X_j$; nothing in the formulation of Cole and Frangakis ensures that $Y_j(x)$ is well-defined if $x \neq X_j$.

One further consequence follows from this refinement of the consistency assumption: we can begin to distinguish between intervention and choice. Many analyses of average causal effects seek to understand the impact of intervening to ensure that all individuals in a population received some particular treatment. Consider the use of observational data in the evaluation of a job training program; suppose the outcome is employment after 3 months and that $Y_j(1, k_1)$ denotes employment status with job training under instructor $k_1 \in K_1$. Assumption ($C1$) would then require that the specific instructor does not affect employment status after 3 months. What then is the meaning of ($C2$)? If $X_j = 1$, then ($C2$) requires that Y_j^{obs}, individual j's employment status if the individual chose to participate in the job training program, is equal to $Y_j(1)$, the individual's employment status if there had been an intervention to ensure the individual participated in the job training program. The values Y_j^{obs} and $Y_j(1)$ need not be equal when $X_j = 1$; for example, resentment arising from compulsion might hinder training. In this setting, the assumption ($C2$) requires that the outcome arising from choosing a treatment, the naturally occurring outcome, is in fact equal to the outcome arising from an intervention to ensure treatment. I would argue that in a number of observational settings this is the real meaning of assumption ($C2$). In some contexts, ($C1$) may hold without ($C2$); similarly, in other contexts, it is also possible that ($C2$) would hold without ($C1$). In any case, the consistency assumption, ($C2$), at least in some contexts, can allow us to distinguish between choice and intervention: the assumption ($C2$) requires that the outcome Y_j^{obs} arising from choice, the outcome naturally arising, is equal to the outcome $Y_j(x)$ arising from intervention. On matters of choice, the econometrics literature on causality[7–9] is certainly ahead of the literature in epidemiology or statistics. However, this may well be due to the fact that exposure to particulate matter, for instance, is not chosen in the same way or with the same intentionality as a job training program.

In light of the above considerations, I propose that ($C1$) be referred to as the assumption of treatment-variation irrelevance (implying for all practical purposes "no multiple-versions-of-treatment") and that if assumption ($C1$) holds, then ($C2$) be referred to as the consistency assumption. Assumption ($C1$) is required not only to state the consistency assumption but also to be able to state the ignorability assumption and to define potential outcomes of the form $Y_j(x)$. Assumption ($C2$), however, relates counterfactual quantities (the potential outcomes) to the data.

Note that if the actual effect of treatment on the treated is of interest, the causal estimand $E[Y|X=1] - E[Y(0)|X=1]$ would only require that $Y(0)$ be well defined that is, ($C1$) would only have to hold for $x = 0$ so that $Y_j(0, k_0) = Y_j(0, k'_0)$ for all $k_0, k'_0 \in K_0$. In some settings this may be more plausible because, as noted above, there will often be far fewer possible variations for a control condition.

THE CONSISTENCY ASSUMPTION AND NONCOMPLIANCE

I would now like to return to the issue of specifying the set K_x. As noted above, it will often be desirable to specify

some set of values K_x for which the consistency assumption is to hold. Cole and Frangakis consider a setting in which x is taken to be "once daily intake of 40 mg of buffered aspirin" and they suggest that k might include components to indicate whether the aspirin is taken in the morning, or whether it is taken with or without food. Suppose we are interested in a treatment comparison of "once daily intake of 40 mg of buffered aspirin taken in the morning" and "no aspirin intake" where "morning" is defined, say, as being between 5:30 AM and 11:00 AM. The consistency assumption and treatment-variation irrelevance assumption for this treatment comparison would then specify K_x to be those values of k_x that indicate the aspirin was taken between 5:30 AM and 11:00 AM and require that that $(C1)$ and $(C2)$ hold for all $k_x \in K_x$, and thus that $Y_j^{obs} = Y_j(x, k_x)$ for all $k_x \in K_x$. Alternatively we might instead be interested in the treatment comparison of "once daily intake of 40 mg of buffered aspirin taken between 7:00 AM and 9:00 AM" and "no aspirin intake." The consistency assumption and treatment-variation irrelevance assumption for this treatment comparison would then specify K_x to be those values of k_x which indicate the aspirin was taken between 7:00 AM and 9:00 AM and require that that $(C1)$ and $(C2)$ hold for all $k_x \in K_x$ and thus that $Y_j^{obs} = Y_j(x, k_x)$ for all $k_x \in K_x$. In any case, a distinction can be drawn between the range of k_x for which the potential outcomes $Y_j(x, k_x)$ can be defined and the range of k_x for which we wish to compare treatments. For a particular treatment comparison, we would require $(C1)$ and $(C2)$ only for values of $k_x \in K_x$ corresponding to the particular treatment comparison of interest. The set K_x would be specified for each x depending on the treatment comparison of interest.

Similarly, we could define a set of individuals J_c such that $j \in J_c$ indicates that individual j in actual fact obtained treatment X_j in a manner inconsistent with the treatment comparison being made; such individuals could be considered as cases of "noncompliance" or "nonadherence." For example, in the treatment comparisons described above, individuals actually taking aspirin at 4:00 PM could be considered as noncompliant or nonadherent to the treatment once daily intake of 40 mg of buffered aspirin taken between 5:30 AM and 11:00 AM. Only individuals j who either do not take aspirin (and are thus in the control group) or who take aspirin between 5:30 AM and 11:00 AM would be included in the set J_c. We might then require that the consistency assumption $(C2)$ hold not for all j but for only for $j \in J_c$. If the extent of treatment variations when treatment is naturally selected is the same as the extent of the treatment variations under interventions then for individual j receiving treatment X_j one could potentially define K_j as the treatment variant for X_j actually received by individual j. We could then potentially define the compliance set J_c by reference to individuals j such that $K_j \in K_{X_j}$. In some settings to address noncompliance we may wish to define the set J_c by principal strata defined by potential outcomes.[10]

When there are individuals who are "noncompliant" with respect to a particular treatment comparison, various techniques for addressing noncompliance could potentially be used in the analysis.[11–14] At least in theory, the same data could be used for several different treatment comparisons. For example, an individual taking aspirin at 10:30 AM would be considered adherent in the first treatment comparison above but not in the second treatment comparison. The same data could thus potentially be used to compare once daily intake of 40 mg of buffered aspirin taken between 5:30 AM and 11:00 AM and "no aspirin intake" and also to compare "once daily intake of 40 mg of buffered aspirin taken between 7:00 AM and 9:00 AM" and no aspirin intake with different groups being considered adherent to the treatment of interest.

STOCHASTIC COUNTERFACTUALS AND VIOLATIONS OF THE TREATMENT-VARIATION IRRELEVANCE ASSUMPTION

The treatment-variation irrelevance assumption and the consistency assumption, like almost all assumptions, are at best approximations; in drawing causal inferences we hope that these are reasonable approximations. Perhaps one further extension will help make these assumptions more reasonable in some contexts. Consider the effect on mortality of driving under the influence of alcohol. One might let k_x indicate the time an individual starts driving under the influence, where x denotes the level of intoxication. If the driving begins at 12:01 AM, there might be an oncoming driver at a particular turn, and driving under the influence would then lead to a crash and so to death; if the intoxicated driving begins at 12:02 AM, the road might be empty and no crash would result. In this case, $(C1)$ will not hold; it will not be the case that $Y_j(x, k_x) = Y_j(x, k_x')$ for all $k_x, k_x' \in K_x$; we would presumably want to include both 12:01 AM and 12:02 AM as possible times in the set K_x. To circumvent this difficulty and the violation of $(C1)$ in this case, we might introduce stochastic counterfactuals[15–18] that is, we might allow Y_j^{obs} and $Y_j(x, k_x)$ to be random variables for each individual j. The outcome obtained both in actuality and under various potential interventions would follow some random distribution. This random distribution would allow for variation in, say, the time when other drivers are on the road. In general, the likelihood of an oncoming driver if the intoxicated driving began at 12:01 AM would likely be approximately the same as the likelihood of an oncoming driver if the intoxicated driving began at 12:02 AM. The assumption of treatment-variation irrelevance, $(C1)$, and the consistency assumption, $(C2)$, could still be formulated as above but the equalities would be equalities in distribution rather than for single values. The distribution of outcomes if intoxicated driving began at 12:01 AM versus at 12:02 AM would be

approximately the same and so the stochastic version of $(C1)$ would likely be reasonable in this case.

Certain aspects of treatment variation, such as which instructor is assigned in a job training program, might also be conceptualized in terms of stochastic counterfactuals. For example, X_j may simply be an indicator of whether individual j participated in the job training program. If there were 3 instructors so that $K_1 = \{1, 2, 3\}$ and if each of the instructors taught any particular trainee with probability 1/3 then the stochastic potential outcome $Y_j(1)$ would take values $Y_j(1, k_1 = 1)$, $Y_j(1, k_1 = 2)$ or $Y_j(1, k_1 = 3)$, each with probability 1/3. In such cases, the stochastic nature of the potential outcomes effectively circumvents the need for $(C1)$, and the consistency assumption for $Y_j(1)$ could then simply be stated as $Y_j(1) = Y_j$ when $X_j = 1$, with $Y_j(1)$ and Y_j both conceived of as random variables. Suppose now that the extent of treatment variations when treatment is naturally selected is the same as the extent of treatment variations under interventions so that one could define K_j as the treatment variant for X_j actually received by individual j, as considered above. In these settings, a quantity such as $E[Y(1)]$ would represent the expected outcome under a stochastic intervention to set each individual's treatment X to 1, by a means k_1 that randomly varies across the population according to how the means is actually randomly determined when treatment is given. In some cases, the probability that $Y_j(x)$ takes a particular value $Y_j(x, k_x)$ may vary with the set of covariates Z. See Taubman et al[19] for a description of data analysis methods for estimating the effects of such hypothetical stochastic interventions.

REFERENCES

1. Cole SR, Frangakis CE. The consistency assumption in causal inference: a definition or an assumption? *Epidemiology*. 2009;20:3–5.
2. Hong G, Raudenbush SW. Evaluating kindergarten retention policy: A case study of causal inference for multilevel observational data. *J Am Stat Assoc*. 2006;101:901–910.
3. Sobel ME. What do randomized studies of housing mobility demonstrate? Causal inference in the face of interference. *J Am Stat Assoc*. 2006;101:1398–1407.
4. Rosenbaum PR. Interference between units in randomized experiments. *J Am Stat Assoc*. 2007;102:191–200.
5. Hudgens MG, Halloran ME. Towards causal inference with interference. *J Am Stat Assoc*. 2008;103:832–842.
6. VanderWeele TJ. Direct and indirect effects for neighborhood-based clustered and longitudinal data. *Sociol Methods Res*. In press.
7. Heckman JJ. The scientific model of causality. *Sociol Methodol*. 2005; 1–98.
8. Heckman JJ, Vytlacil EJ. Econometric evaluation of social programs. Part II: Using the marginal treatment effect to organize alternative economic estimators to evaluate social programs and to forecast their effects in new environments. In: Heckman J., Leamer E, eds. *Handbook of Econometrics*. Vol. 6B. Amsterdam: Elsevier; 2007:4875–5144.
9. Heckman JJ. Econometric causality. *Int Stat Rev*. 2008;76:1–27.
10. Frangakis CE, Rubin DB. Principal stratification in causal inference. *Biometrics*. 2002;58:21–29.
11. Goetghebeur E, Molenberghs G. Causal inference in a placebo-controlled clinical trial with binary outcome and ordered compliance. *J Am Stat Assoc*. 1996;91:928–934.
12. Imbens GW, Rubin DB. Bayesian inference for causal effects in randomized experiments with noncompliance. *Ann Stat*. 1997;25:305–327.
13. Robins JM. Correction for non-compliance in equivalence trials. *Stat Med*. 1998;17:269–302.
14. Robins JM, Finkelstein D. Correcting for non-compliance and dependent censoring in an AIDS clinical trial with inverse probability of censoring weighted (IPCW) Log-rank tests. *Biometrics*. 2000;56:779–788.
15. Greenland S. Interpretation and choice of effect measures in epidemiologic analyses. *Am J Epidemiol*. 1987;125:761–768.
16. Robins JM, Greenland S. The probability of causation under a stochastic model for individual risk. *Biometrics*. 1989;45:1125–1138.
17. Robins JM, Greenland S. Comment on: "Causal inference without counterfactuals" by AP Dawid. *J Am Stat Assoc*. 2000;95:477–482.
18. van der Laan MJ, Haight TJ, Tager IB. Response to "Hypothetical interventions to define causal effects" by MA Hernán. *Am J Epidemiol*. 2005;162:621–622.
19. Taubman SL, Robins JM, Mittleman MA, Hernán MA. Alternative approaches to estimating the effects of hypothetical interventions. In: Proceedings of the 2008 Joint Statistical Meeting; 2008; Alexandria, VA.