Collection of Biostatistics Research Archive
COBRA Preprint Series

Year 2011 *Paper* 77

Causal inference under multiple versions of treatment

Tyler J. VanderWeele* Miguel A. Hernan†

*Harvard University, tvanderw@hsph.harvard.edu
†Harvard University, mhernan@hsph.harvard.edu

Causal inference under multiple versions of treatment

Tyler J. VanderWeele and Miguel A. Hernan

Abstract

In this article we discuss the no-multiple-versions-of-treatment assumption and extend the potential outcomes framework to accommodate causal inference under violations of this assumption. A variety of examples are discussed in which the assumption may be violated. Identification results are provided for the overall treatment effect and the effect of treatment on the treated when multiple versions of treatment are present and also for the causal effect comparing a version of one treatment to some other version of the same or a different treatment. Further identification and interpretative results are given for cases in which a treatment variable is dichotomized to create a new treatment variable for which there are effectively "multiple versions" and also for effects defined by setting the version of treatment to a prespecified distribution. Some of the identification results bear resemblance to identification results in the literature on direct and indirect effects. We describe some settings in which ignoring multiple versions of treatment, even when present, will not lead to incorrect inferences.

1. Introduction

The potential outcomes framework for causal inference employs a number of assumptions (Neyman, 1923; Rubin, 1974, 1990; Robins, 1986). One of the assumptions that is generally made, either implicitly or explicitly, is an assumption that is sometimes described as the "no-multiple-versions-of-treatment assumption"; the assumption is part of what Rubin defined as the "Stable Unit Treatment Value Assumption" or SUTVA (Rubin, 1980, 1986). The assumption is made so that the potential outcomes for each individual under each possible treatment are well defined and take on a single value. If there are multiple versions of treatment present, as might arise for surgery treatment say if there are different surgeons who perform the surgery, and if these different versions of treatment give rise to different potential outcomes, then this assumption will be violated.

To circumvent the problems created by multiple versions of treatment in such contexts, one might restrict inference to a single version of treatment or, more generally, redefine each version of treatment as a different treatment. For example, one could consider the effect of surgery conducted by each particular surgeon rather that the effect of surgery generally. Such redefinition of the treatment of interest would make the no-multiple-versions-of-treatment assumption more reasonable. However, redefining each version of treatment as a different treatment may not always be possible or desirable. One may not have data on which version each patient received. Moreover, if a patient needs to decide whether or not to undergo surgery but has no control over the choice of the surgeon who will actually perform the surgery, the average effect of the surgery treatment generally, rather than the average effect of surgery for each particular surgeon, may be what is most relevant. Such average effects of a particular surgical procedure (averaged over the surgeon who administers it) may also be of interest from a policy perspective. If, for example, in the treatment of cancer patients we are comparing the effects of radiation versus surgery, although different surgeons (i.e. different versions of treatment) may have different effects on survival, from a policy perspective, we could not simply select the most competent surgeon to perform all of the surgeries as the number of surgeries needed would be far too numerous for one surgeon to undertake. The policy question of interest here would be evaluating the overall survival rates of radiation versus surgery, taking into account the fact that not all surgeons are equally skilled.

Motivated by the above considerations, the purpose of this article is to consider causal inference under violations of the no-multiple-versions-of-treatment assumption. The remainder of the paper is organized as follows. In section 2 we review the potential outcomes framework and discuss how it can accommodate settings of multiple versions of treatment; we discuss the definition of causal effects under multiple versions of treatment. In section 3, we discuss identification of these effects when the first treatment is assigned and then a particular version of treatment is assigned. We discuss the interpretation of causal effect estimates under multiple versions and we consider when multiple versions of treatment can be ignored. In section 4, we will consider what new questions might be of substantive interest when multiple versions of treatment are present ; some of the identification results bear certain resemblances to the analysis of direct and indirect effects (Robins and Greenland, 1992; Pearl, 2001; Geneletti, 2007), though the precise technical details are distinct. In section 5, we discuss cases in which the ordering is version then treatment rather than treatment then version. An illustration is given in section 6. In section 7, we offer some concluding remark and discuss how the extensions given in this paper to allow for violations of the no-multiple-versions-of-treatment assumption parallels in certain ways extensions described elsewhere for violations of the other major component of SUTVA, the no-interference assumption (Sobel, 2006; Hong and Raudenbush, 2006; Rosenbaum, 2007; Hudgens and Halloran, 2008; Tchetgen Tchetgen and VanderWeele, 2010).

2. Potential Outcomes and the No-Multiple-Versions-of-Treatment Assumption

We will use $j = 1, ..., N$ to index the individuals in the population. Let A_j and Y_j denote respectively the actual treatment received by and the actual outcome for individual j. Under the standard potential outcomes framework (Rubin, 1974, 1990), one might use $Y_j(a)$ to denote the potential outcome Y for individual j if treatment A were set, possibly contrary to fact, to the value a. Suppose treatment takes values in some set $\mathcal{A}$; often $\mathcal{A} = \{0, 1\}$ with 0 indicating the control condition and 1 indicating the treatment condition. Articulating the potential outcomes framework in this way requires what Rubin called the "Stable Unit Treatment Value Assumption" or "SUTVA." As Rubin (1980) points out that notation such as $Y_j(a)$ effectively presupposes (i) that if individual j is given treatment a then individual j's outcome under treatment a does not depend on which treatment individual $j' \neq j$ received and (ii) that there do not exist multiple versions of treatment a which might give rise to different outcomes depending on which version is administered. The first of these assumptions is sometimes referred to as "no-interference" which Rubin (1980) attributes to Cox (1958); the second assumption is a "no-versions-of-treatment assumption" which Rubin attributes to Neyman (1935). Included also within SUTVA is an assumption which in other literature is sometimes referred to as consistency. The consistency assumption (Robins, 1986) states that $Y_j(a) = Y_j$ when $A_j = a$ i.e. that the value of Y which would have been observed if A had been set to what it in fact was is equal to the value of Y which was in fact observed. The consistency assumption ties the potential outcomes (or counterfactual data) to the observed data. Under Rubin's articulation of SUTVA, if there is only one version of treatment, then if $A_j = a$, the manner in which treatment A_j was in fact set to a is irrelevant, so $Y_j(a)$ is well defined and is equal to Y_j when $A_j = a$. Rubin's SUTVA thus includes a no-multiple-versions-of-treatment assumption and this no-multiple-versions-of-treatment assumption itself includes the consistency assumption.

The assumption of "no multiple versions of treatment" and SUTVA generally are relevant both to experimental and non-experimental studies. Although SUTVA is often only explicitly noted in non-experimental observational research, the assumption is important in the interpretation of causal effects even in randomized trials.

As noted above, one potential approach for handling multiple versions of treatment would be to redefine the treatment variable A so as to include the version of treatment. This then generates an expanded set of potential outcomes one for each "treatment level" (defined by the version of treatment) and under this redefined treatment the no-multiple-versions-of-treatment assumption will hold; limited sample size for each version may limit the effectiveness of this approach. Moreover, if we did not observe the version, redefining treatment so as to indicate version makes it difficult to identify causal effects since we would then not be observing what was redefined to be the treatment variable. Schafer and Kang (2009), however, have presented work that makes some progress with this approach; their approach does still assume that some indicators related to version are at least available.

Rather than redefining treatment in this way, additional progress can be made by instead introducing separate notation for the treatment itself and for the version of treatment. By taking this approach instead we will sometimes be able to define and identify causal effects even if we do not observe the version of treatment that each individual received. We follow Cole and Frangakis (2009) and VanderWeele (2009) to extend the potential outcomes notation to allow for multiple versions of treatment. Let $Y_j(a, k^a)$ be the potential outcome for individual j if treatment A is set to value a by means k^a where k^a takes values in some set $\mathcal{K}^a = \{1, ..., n^a\}$. For example, if comparing surgery at a hospital ($A = 1$) to a control condition ($A = 0$), the set $\mathcal{K}^0$ may simply be a singleton if there were only one version of the control condition and $\mathcal{K}^1$ might be the set $\{1, 2, 3\}$ indicating surgeon 1, 2 or 3 respectively and the potential outcomes $Y_j(1, 1)$, $Y_j(1, 2)$ and $Y_j(1, 3)$ would indicate how individual j would fare under surgery by surgeon 1, 2 or 3 respectively. For the next two sections we focus on settings in which causal ordering of variables is treatment then version (rather than version then treatment). In the surgery example, an individual is first assigned to surgery then to a surgeon.

If the two treatments being compared were surgery and radiation, some aspects of treatment variation (e.g. specific hospital, time of treatment initiation) may be common to the treatments being compared but generally not all will be. For each treatment a, we will consider a distinct set of versions $\mathcal{K}^a = \{1, ..., n^a\}$. For individuals with $A_j = a$ we let K_j^a denote the version of treatment $A_j = a$ actually received by individual j; for individuals with $A_j \neq a$ we define $K_j^a = 0$ so that $K_j^a \in \{0\} \cup \mathcal{K}^a$. For notational convenience, we define the vector $\mathbf{K}_j = (K_j^a : a \in \mathcal{A})$ and recall that $K_j^a = 0$ for $a \neq A_j$ so that $\mathbf{K}_j$ thus denotes a vector in which all of the entries are 0 except the entry corresponding to the treatment that individual j actually received and this entry indicates what version of that treatment was in fact received by individual j. Note also that $\mathbf{K}_j$ gives no more information than A_j and $K_j^{A_j}$ together. Note also that there is not variation independence between A_j and $\mathbf{K}_j$; once we know the vector $\mathbf{K}_j$, we know A_j; but A_j does not uniquely determine $\mathbf{K}_j$.

Under this expanded potential outcomes notation the no-multiple-versions-of-treatment assumption can then simply be articulated as that

$$Y_j(a, k^a) = Y_j(a, k'^a) = Y_j(a, \bullet) \text{ for all } j, a, \text{ and } k^a, k'^a \in \mathcal{K}^a. \tag{1}$$

If (1) holds then the consistency assumption is simply that for all j,

$$\text{if } A_j = a \text{ then } Y_j = Y_j(a, \bullet) \tag{2}$$

and (1) and (2) together would bring us back to Rubin's articulation of the no-multiple-versions-of-treatment assumption. VanderWeele (2009) referred to (1) as an assumption of treatment variation irrelevance (which need not necessarily imply consistency assumption (2)). Under multiple versions of treatment, if the version has no relevance to the outcome under consideration then (1) will hold and, for all practical purposes, there are "no multiple versions of treatment," at least with regard to the outcome Y under consideration. Note, however, that the version of treatment may be irrelevant (i.e. (1) holds) for some outcome but may not be irrelevant for a different outcome.

If the treatment variation irrelevance assumption (1) is violated we may still articulate a consistency assumption as follows. The consistency assumption would then require for all j,

$$Y_j = Y_j(a, k^a) \text{ when } A_j = a \text{ and } K_j^a = k^a. \tag{3}$$

This expanded potential outcomes notation essentially presupposes that, for a subject with $A_j = a$ and $K_j^a = k^a \in \mathcal{K}^a$, (i) the potential outcomes $Y_j(a, k'^a)$ with $k'^a \neq k^a$ are well defined, and (ii) the potential outcomes $Y_j(a^*, k^{a^*})$ with $a^* \neq a, k^{a^*} \in \mathcal{K}^{a^*}$ are well defined. In the surgery example, for an individual who in fact received surgery by surgeon 1, we could conceive of what would have happened to this individual had they received surgery from surgeon 2 or surgeon 3, and also what would have happened if surgery had not been given at all; for an individual who did not receive surgery we could conceive of what would have happened to the individual had the individual received surgery from surgeons 1, 2 or 3.

The average causal effect comparing treatment a, version k^a with treatment a^*, version k^{a^*} is defined by:

$$\mathrm{E}\{Y(a, k^a)\} - \mathrm{E}\{Y(a^*, k^{a^*})\}. \tag{4}$$

The potential outcomes $Y(a, k^a)$ might be conceived of as consisting of joint interventions on both A_j and K_j^a to respectively set them to levels a and k^a (Pearl and Robins, 1995; Pearl, 2001). If the potential outcomes $Y(a, k^a)$ or $Y(a^*, k^{a^*})$ are only defined for individuals for whom S takes certain levels of some covariate set S then we may instead be interested in the conditional causal effect:

$$\mathrm{E}\{Y(a, k^a)|S = s\} - \mathrm{E}\{Y(a^*, k^{a^*})|S = s\}. \tag{5}$$

Note that a special case of (4) and (5) would be when the treatment a is the same and only different versions, k^a and k'^a are being compared.

A well known causal effect is the effect of treatment on the treated, which is usually represented as $\mathrm{E}\{Y(a)|A = a\} - \mathrm{E}\{Y(0)|A = a\}$. In the presence of multiple versions of treatment $A = a$, the counterfactual outcome $Y(a)$ can be represented as $Y(a, K^a)$. For individuals for whom $A = a$, K^a is simply the version of treatment that was actually received. Note that K^a is a random variable whose value may be different for different subjects, not a fixed value k^a. Suppose now that there is only one version of treatment for the control condition, $A = 0$, so that $K^0 = \{1\}$ and the only potential outcome for each individual under the control condition is $Y_j(0) \equiv Y_j(0, 1)$. Provided $Y_j(0)$ is well-defined for individuals with treatment $A_j = a \neq 0$, we could then define the effect of treatment on the treated as:

$$\mathrm{E}\{Y(a, K^a)|A = a\} - \mathrm{E}\{Y(0)|A = a\}. \tag{6}$$

We will use additional notation to define another familiar causal effect, the overall treatment effect, but here in the setting with multiple versions of treatment. For individuals with $A_j = a$ we defined K_j^a to be the version of treatment $A_j = a$ actually received by individual j. In some cases, for individuals with $A_j = a$, we might be willing to conceive of a counterfactual variable $K_j^{a^*}(a^*)$, $a^* \neq a$, corresponding to the version of treatment a^* that an individual would have received had they in fact been assigned to treatment a^* rather than a. For example, in the surgery context, for an individual j who did not receive surgery ($A_j = 0$), the variable $K_j^1(1)$ would denote which surgeon individual j would have been assigned to had the individual in fact undertaken surgery i.e. whether the individual would have been assigned to surgeon 1, 2 or 3. We then assume a consistency assumption for K_j^a, namely, $K_j^a = K_j^a(a)$ when $A_j = a$. For each individual j, we assume there is a fixed version that would have been received had the individual been given treatment a^*; as with the case of stochastic counterfactuals (Robins and Greenland, 2000), analogous results to those that follow would hold if these "counterfactual versions of treatment" were assumed stochastic.

The variable $K_j^{a^*}(a^*)$, $a^* \neq A_j$ bears some resemblance to the counterfactual value of the mediator in the literature on direct and indirect effects but, unlike in the mediation context, counterfactuals of the form $Y_j(a, K_j^{a^*}(a^*))$ are only defined when a and a^* coincide; that is, if a^* is the actual (or the counterfactual) treatment the only possible versions of treatment are different versions $K^{a^*}$of treatment a^*. If we are willing to postulate variables $K_j^{a^*}(a^*)$, $a^* \neq A_j$ then we can define the overall treatment effect comparing giving everyone treatment a versus treatment a^* by

$$\mathrm{E}\{Y(a, K^a(a))\} - \mathrm{E}\{Y(a^*, K^{a^*}(a^*))\}. \tag{7}$$

Assuming the relevant counterfactuals exist we can define, $Y(a) \equiv Y(a, K^a(a))$ and the expression above is simply $\mathrm{E}\{Y(a)\} - \mathrm{E}\{Y(a^*)\}$. In the following section we will discuss the identification of these various treatment effects. Later we will also consider the definition and identification of some additional causal effect measures.

3. Identification of Causal Effects Under Multiple Versions of Treatment

We now consider identification in the setting of multiple versions of treatment. We partition the set of covariate (potential confounders) into two sets. We let W indicate a set of covariates that may be causes of treatment A or may be, for one or more treatment levels a, causes of which version of treatment K^a is administered but are not causes of Y except through either treatment or the version of treatment; let C denote all other covariates. We partition the covariates in this way because as we will see below, to identify certain causal effect we do not need data on W i.e. certain effects, but not others, would still be identified even if data on W were unavailable. We can then represent the relations between treatment, version, outcome and covariates as in Figure 1.

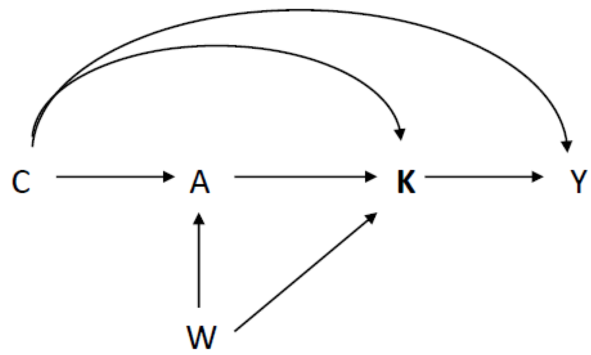


Figure 1. Causal diagram illustrating relationships between treatment A, version **K**, outcome Y and confounding variables C and W.

Note that because $\mathbf{K}$ contains all of the information in A, there is no arrow directed from A to Y. An example of a covariate in W in the context of the surgery example might be the particular health plan that an individual has that might affect both the probability that the individual receives surgery and, if the individual does receive surgery, also which particular surgeon conducts the operation, but would perhaps not affect the outcome except through whether surgery is received and who performs the surgery.

The results below will require that the covariate sets C and W are sufficiently rich so that certain no unmeasured confounding assumptions (also sometimes referred to as "exchangeability" or "ignorability" assumptions) are met. We will use the notation $A \amalg B|C$ to denote that A is independent of B given C. We will first consider the following assumption: whether we have

$$Y(a, k^a) \amalg \{A, \mathbf{K}\}|C \text{ for all } a \in \mathcal{A}, k^a \in \mathcal{K}^a \tag{8}$$

In other words, we will consider whether, within strata of the covariates C, groups defined by treatment and version are, for all $a \in \mathcal{A}, k^a \in \mathcal{K}^a$, comparable in their potential outcomes under treatment a, version k^a. Note that assumption (8) would hold if treatment A were randomized (or randomized conditional on C) and if, conditional on treatment $A = a$ (or conditional on $\{A = a, C = c\}$), version of treatment K^a were also randomized. Intuitively, (8) states that the set C suffices to control for confounding of the joint effect of treatment and version on the outcome. Assumption (8) will hold if Figure 1 represents a causal directed acyclic graph (Pearl, 2009). Note that in Figure 1, C blocks all backdoor paths from the set $\{A, \mathbf{K}\}$ to Y. Note also that since $\mathbf{K}$ in fact contains all the information in A, assumption (8) could also be written as $Y(a, k^a) \amalg \mathbf{K}|C$; the two are equivalent.

We then have the following identification results. The proofs of all results are given in the online supplementary materials; some of the proofs bear resemblance to certain identification results in the literature on direct and indirect effects.

Theorem 1. If $Y_j(a, k^a)$ and $Y_j(a^*, k^{a^*})$ are well defined for all individuals j and if the no-unmeasured-confounding assumption (8) holds then

$$\begin{aligned}\mathrm{E}\{Y(a,k^a)\} - \mathrm{E}\{Y(a^*,k^{a^*})\} &= \textstyle\sum_c \mathrm{E}\{Y|A=a,K^a=k^a,C=c\}pr(c) \\ &\quad - \textstyle\sum_c \mathrm{E}\{Y|A=a^*,K^{a^*}=k^{a^*},C=c\}pr(c).\end{aligned}$$

Theorem 1 thus allows for the identification of the average causal effect comparing treatment a, version k^a with treatment treatment a^*, version k^{a^*}. We could control for W as well in assumption (8) and Theorem 1, but data on W is not necessary to identify the effect given in Theorem 1; data on W will, however, be needed for the overall treatment effect result given below. The proof of Theorem 1 given in the online supplementary materials is for the identification of $\mathrm{E}\{Y(a,k^a)\}$; this proof is completely isomorphic to the proofs often used for the identification of counterfactuals for controlled direct effects (Robins and Greenland, 1992; Pearl, 2001). Note however that the effect itself is somewhat different; even in the special case with $a = a^*$ but $k^a \neq k'^a$, the effect $\mathrm{E}\{Y(a,k^a)\} - \mathrm{E}\{Y(a,k'^a)\}$ corresponds to having the "treatment" variable fixed not the "mediator" variable as in controlled direct effects. If $Y(a,k^a)$ and/or $Y(a^*,k^{a^*})$ are only defined for those with certain covariate values of $S \subseteq C$ then the result in Theorem 1 can be made conditional on $S = s$. In the online supplement we also discuss settings in which there may be an effect, Q, of treatment A that affects both version $\mathbf{K}$ and the outcome Y, a setting sometimes referred to as "time-dependent" confounding. Analogous results hold but identification formulas are different.

We now consider the identification of the effect of treatment on the treated.

Theorem 2. If there is only one version of treatment for the control condition so that and $Y_j(0)$ is well defined for all individuals j and if $Y(0) \amalg \{A, K^0\}|C$ then

$$\mathrm{E}\{Y(a,K^a)|A=a\} - \mathrm{E}\{Y(0)|A=a\} = \mathrm{E}(Y|A=a) - \textstyle\sum_c \mathrm{E}\{Y|A=0,C=c\}pr(c|A=a).$$

Note that $Y(0) \amalg \{A, K^0\}|C$ would hold under assumption (8) above; however, for the application of Theorem 2 we only need the weaker condition $Y(0) \amalg \{A, K^0\}|C$ rather than that (8) hold for all $a \in \mathcal{A}, k^a \in \mathcal{K}^a$. Once again with Theorem 2 we do not need to control for the $A - \mathbf{K}$ confounders, W.

If we are willing to postulate variables $K_j^{a^*}(a^*)$, $a^* \neq A_j$, so that we can define $Y(a) = Y(a, K^a(a))$ we can also consider the identification of the overall treatment effect. As will be shown in Theorem 3, overall treatment effects are identified if

$$Y(a) \amalg A|\{C, W\} \text{ for all } a \tag{9}$$

Assumption (9) requires that, within strata of the covariates $\{C, W\}$, groups defined by treatment are comparable in their potential outcomes $Y(a) = Y(a, K^a(a))$. Assumption (9) would hold if treatment A were randomized (or randomized conditional on $\{C, W\}$); assumption (9) holds if Figure 1 is a causal directed acyclic graph (Pearl, 2009).

Theorem 3. If $K_j^a(a)$, $K_j^{a^*}(a^*)$, $Y_j(a, K_j^a(a))$ and $Y_j(a^*, K_j^{a^*}(a^*))$ are well defined for all individuals j and if (9) holds then $Y(a) - Y(a^*) =$

$$\begin{aligned} \mathrm{E}\{Y(a, K^a(a))\} - \mathrm{E}\{Y(a^*, K^{a^*}(a^*))\} &= \sum\nolimits_c \mathrm{E}\{Y|A = a, C = c, W = w\}pr(c, w) \\ &\quad - \sum\nolimits_c \mathrm{E}\{Y|A = a^*, C = c, W = w\}pr(c, w). \end{aligned}$$

Note for the quantity on the right hand side of the equation in Theorem 3 to itself be estimable from data we would need "positivity" (or "experimental treatment assignment" assumption) to hold for both $\{C, W\}$, not just C i.e. $0 < P(A = a|C = c, W = w) < 1$ for all a, c and w.

Theorem 3 states that even under violations of the no-multiple-versions-of-treatment assumption we can use the ordinary identification formula for overall treatment effects but control needs to be made not simply for variables C that might confound the relationship between treatment assignment and outcome but also for variables W that may affect both treatment assignment and version of treatment (even if these variables do not also affect the outcome except through treatment or version of treatment). Control must be made for these variables essentially because for the overall treatment effect we are examining the effect of A on Y and W is a confounder of the relationship between A and Y as it affect both A and also Y through $\mathbf{K}$. In Figure 1, if control is not made for W then there is a backdoor path from A to Y, namely, $A \leftarrow W \rightarrow \mathbf{K} \rightarrow Y$. In the context of multiple versions of treatment a sufficient set of confounder for the effect of A on Y would need to include W. In the online supplementary materials we give a numerical example showing that without controlling for a common cause of A and $\mathbf{K}$, one can obtain biased estimates of the overall treatment effect. Note that if treatment is randomized there will be no common causes of treatment and version.

Several observations emerge from the results above. First, Theorems 2 and 3 demonstrate what may be intuitively clear, that it is not necessary to have data on the versions of treatment in order to estimate the effect of treatment on the treated or the overall treatment effect. For the overall treatment effect, adjustment needs to be made not only for common causes of treatment and outcome but also common causes of treatment and version of treatment. For the effect of treatment on the treated, if there are no multiple versions of treatment for the control condition, it is not necessary to have data on common causes of treatment and version. Finally, as shown in Theorem 1, if data are available on the version of treatment then one can identify causal effects comparing a version of one treatment to some other version of the same or a different treatment and to identify such effects one again does not need data on the common causes of treatment and version.

Theorem 3, in particular, has important implications for settings in which the multiple versions of treatment assumption is violated and the violation is ignored. The result implies that the ordinary estimator for average causal effects can be interpreted as a contrast between (i) the average outcome that would be expected if everyone had been assigned treatment $A = 1$ with each individual receiving the version that would have been received had they been assigned $A = 1$ versus (ii) the average outcome that would be expected if everyone had been assigned treatment $A = 0$ with each individual receiving the version that would have been received had they been assigned $A = 0$. The estimate carries this interpretation provided adjustment is made for all confounders of the relationship between treatment A and outcome Y; importantly, however, within the context of multiple versions of treatment, these confounders include common causes of treatment and version. Data on version is not needed, the multiple-versions-of-treatment assumption does not need to hold, but data on common causes of treatment and version are needed to interpret the estimate as a causal effect.

One way to think about these results are that if we conceive of a treatment and version combination, (a, k^a), as a "regime", and a treatment a, along with an unknown rule relating treatment to version as a "policy," then Theorem 1 states that if the usual conditional independence assumption holds at the level of the regime (assumption 8) we can identify causal effects at the level of the regime. Theorem 3 states that if the usual conditional independence assumption holds at the level of the policy (assumption 9), then for the set of policies in place in the study, we can identify average causal effects at the level of the policy.

4. New Causal Effects and Applications with Multiple Versions of Treatment

In this section we will consider the identification and interpretation of a different type of causal effect which arises by setting the version of treatment to various prespecified distributions. These prespecified distributions may be fixed or may be defined by those of certain treatment groups or by those of individuals with certain pretreatment covariate values. For the results in this section we will rely principally on assumption (8) above but for one result we will consider another "exchangeability" or "no unmeasured confounding" condition, namely whether

$$K^a(a) \amalg A|\{C, W\} \text{ for all } a \tag{10}$$

In other words, we will consider whether, within strata of the covariates $\{C, W\}$, groups defined by treatment are comparable in the versions of treatment they would have received under each possible treatment a. Note that assumption (10) would hold if treatment A were randomized (or randomized conditional on $\{C, W\}$. Intuitively, (10) states that within strata of $\{C, W\}$ the version of treatment which an individual would be assigned if given treatment a is independent of the treatment actually received; note the set $\{C, W\}$ blocks all backdoor paths from A to $\mathbf{K}$; assumption (10) would thus hold if Figure 1 were a causal directed acyclic graph (Pearl, 2009).

Taubman et al. (2008) considered what the incidence of coronary heart disease would be if everyone exercised at least 30 minutes per day compared to what it actually was, $\mathrm{E}[Y]$. Note that there are clearly multiple versions of treatment for both $A = 1$ ("exercising at least 30 minutes per day") and $A = 0$ ("exercising less than 30 minutes a day"). With slight abuse of notation (by not beginning the indices of each K^a with 1), we might index K^1 by $\{30, 31, 32...\}$ and K^0 by $\{0, 1, 2, ...29\}$. Taubman et al. (2008) considered two hypothetical intervention regimes g that would ensure "exercising at least 30 minutes per day." Here we provide formal identification results for the two hypothetical intervention regimes that Taubman et al. (2008) discussed. Under the first intervention regime, those with $A = 1$ who in fact exercised at least 30 minutes were allowed to retain their actual number of minutes of exercise K^1 and those with $A = 0$ who in fact exercised less than 30 minutes were, under treatment, assigned version of treatment $K^1 = 30$. The counterfactual quantity of interest was thus $\mathrm{E}[Y(1, G)]$ where $G = K^1$ if

$A = 1$ and $G = 30$ otherwise. Taubman et al. described this as a "threshold intervention." Since, $\mathrm{E}\{Y(1,G)|A=1\} = \mathrm{E}\{Y(1,K^1)|A=1\} = \mathrm{E}(Y|A=1)$, to identify $\mathrm{E}\{Y(1,G)\}$ it suffices to identify $\mathrm{E}\{Y(1,G)|A=0\} = \mathrm{E}\{Y(1,30)|A=0\}$. Identification conditions for this quantity are given in the following result.

Theorem 4. If for some $a^* \neq a$, and some k^a, $Y_j(a,k^a)$ is well defined for all individuals with $A_j = a^*$and if the no-unmeasured-confounding assumption (8) holds then

$$\mathrm{E}\{Y(a,k^a)|A=a^*\} = \sum\nolimits_c \mathrm{E}(Y|a,k^a,c)pr(c|a^*).$$

The second regime g considered by Taubman et al. again lets individuals with $A = 1$ who in fact exercised at least 30 minutes retain their actual number of minutes of exercise K^1; under the second regime those with $A = 0$ who in fact exercised less than 30 minutes were, under treatment, randomly assigned a version of treatment K^1 from the distribution of those with $A = 1$ who had the same covariates. Taubman et al. described this second regime as a representative regime. If we now let G_j denote a randomly assigned version, K^1, of treatment from the distribution of those with $A = 1$ with covariates C_j then the quantity $\mathrm{E}\{Y(1,G)|A=0\}$ is needed to identify the counterfactual incidence of coronary heart disease under the "representative regime." This quantity is identified by the following result.

Theorem 5. For individuals j with $A_j = a^* \neq a$, let G_j^a be a random variable with distribution defined by $pr(K^a = k^a|A = a, C = C_j)$. If for all j such that $A_j = a^*$, the potential outcome $Y_j(a,k^a)$ is well defined for all $k^a \in supp(G_j^a)$ and if the no-unmeasured-confounding assumption (8) holds then

$$\mathrm{E}\{Y(a,G^a)|A=a^*\} = \sum\nolimits_{c,k^a} \mathrm{E}(Y|a,k^a,c)pr(K^a = k^a|a,c)pr(c|a^*). \tag{11}$$

Note that the quantity $\mathrm{E}\{Y(a,G^a)|A=a^*\}$ in Theorem 5 does not necessary reflect what would happen if we were to set the version of treatment to the version it would have been if those with $A = a^*$ had in fact been given treatment $A = a$. The formula in (11) will however identify this latter quantity if, as stated formally in the next theorem, for individuals with $A_j = a^*$, $K_j^a(a)$ is well defined and if the set of covariates for which control is made contains both C and W in Figure 1 so that in addition to identification assumption (8), identification assumption (10) also holds. Whereas Theorem 5 gave a result for randomly setting version of treatment to the distribution of those with $A = a$, Theorem 6 gives a result for randomly setting version of treatment to the distribution of those with $A = a^*$ had they been given treatment $A = a$.

Theorem 6. If for all individuals j with $A_j = a^* \neq a$, the potential outcome $K_j^a(a)$ is well defined then let $G_j^{a|a^*}$ be a random variable with distribution defined by $pr(K^a(a) = k^a|A = a^*, C = C_j, W = W_j)$. If for all j such that $A_j = a^*$, the potential outcome $Y_j(a,k^a)$ is well defined for all $k^a \in supp(G_j^{a|a^*})$ and if the no-unmeasured-confounding assumptions (8) and (10) hold then

$$E\{Y(a, G_j^{a|a^*})|A = a^*\} = \sum_{c,w,k^a} E\{Y|a, k^a, c, w\} pr(K^a = k^a|a, c, w) pr(c, w|a^*).$$

Under assumptions (8) and (10) of Theorem 6, ordinary estimators of treatment effects will, for each treatment group, estimate the effect of randomly setting the version to one selected from the distribution of versions of those who were in fact in that treatment group; note this is a somewhat stronger interpretation than that provided in Theorem 3. The analytic formulas in Theorems 5 and 6 bear resemblances to those for so-called "natural direct and indirect effects" (Robins and Greenland, 1992; Pearl, 2001) when these effects are identified. However, unlike in the literature on natural direct effects, all of the assumptions made here, namely (8)-(10), would be satisfied if both treatment and version were randomized whereas the identification of natural direct and indirect effects requires counterfactual independence assumptions that may not hold even in a doubly randomized trial (Robins, 2003).

We define and provide an identification result for one further counterfactual quantity. In health disparities research, health outcomes are compared by strata of race or socioeconomic status. In some cases, access to care or receipt of a treatment or procedure may be equal across strata of racial groups but health outcome disparities may still persist. One possibility is that race itself modifies the effect of treatment. Another possibility is that there may in fact be disparities in the version of the treatment being administered. Thus even if outcome disparities are not explained by disparities in the receipt of treatment, they may be explained by disparities in the version of treatment. Let A denote some treatment; for example, for patients with acute respiratory distress syndrome, $A = 1$ might denote low-volume ventilation and $A = 0$ traditional ventilation (Acute Respiratory Distress Syndrome Network, 2000). Let K^1 denote the version of treatment $A = 1$ (e.g. K^1 might denote the quality of the monitoring for low-volume ventilation) and let Y denote the health outcome (e.g. 180 day survival). Let $S \subseteq C$ denote one or more covariates of interest; here we will let S denote race. We might, for example, then be interested in how much better outcomes would have been for black individuals ($S = 1$) who received treatment if they had obtained the same quality of treatment as white individuals ($S = 0$). If for individuals with $S = 1$, we let G_j denote a randomly assigned version of treatment K^1 from the distribution of those with $A = 1$, $S = 0$ and with covariates $C \backslash S$ equal to $C_j \backslash S_j$ then this quantity is given by

$$\begin{aligned} & E\{Y(1, G)|A = 1, S = 1\} - E\{Y(1, K^1)|A = 1, S = 1\} \\ = \; & E\{Y(1, G)|A = 1, S = 1\} - E(Y|A = 1, S = 1) \end{aligned}$$

The counterfactual quantity $E\{Y(1, G)|A = 1, S = 1\}$ is identified by the following result.

Theorem 7. Let $S \subseteq C$ and let $G_j^{a,s'}$ be a random variable with distribution defined by $pr(K^a = k^a|A = a, S = s', C \backslash S = C_j \backslash S_j)$. If for all j such that $A_j = a$ and $S = s$, the potential outcome $Y_j(a, k^a)$ is well defined for all $k^a \in supp(G_j^{a,s'})$ and if the no-unmeasured-confounding assumption (8) holds then

$$\mathrm{E}\{Y(a, G^{a,s'})|A=a, S=s\} = \sum_{c,k^a} \mathrm{E}(Y|a, k^a, c, s)pr(K^a = k^a|a, c, s')pr(c|a, s).$$

Note that the hypothesis that different versions of treatment were given to groups $S = 1$ and $S = 0$, not because of discrimination, but because differing versions of treatment, k^a and k'^a, have differing effects across strata of S could be examined by contrasting $\mathrm{E}\{Y(a, k^a)|c, S = 1\} - \mathrm{E}\{Y(a, k'^a)|c, S = 1\}$ and $\mathrm{E}\{Y(a, k^a)|c, S = 0\} - \mathrm{E}\{Y(a, k'^a)|c, S = 0\}$ which would be identified under assumption (8).

5. When Version Precedes Treatment and Consequences of Dichotomization

All the material thus far considered a setting where treatment is set first and then the version of treatment; in this section we consider the reverse scenario. For example, often researchers will dichotomize or otherwise categorize or coarsen a continuous exposure to simplify an analysis. For example, if the continuous exposure is the number of minutes of exercise, a researcher may form a new dichotomous "treatment" variable defined by exercising at least 30 minutes. One might then speak of different "versions" of the treatment "exercise at least 30 minutes" e.g. exercise 30 minutes, exercise 31 minutes, etc. In the analysis for multiple versions of treatment given above we have presupposed that the causal order of the variables was treatment then version. However, when a continuous exposure has been dichotomized an alternative conceptualization might be that the version then treatment.

Suppose now, in contrast to the analysis in the previous two sections, that K is a treatment variable such that the support of K is of cardinality greater than 2. Suppose also we partition the support of K into two sets V_0 and V_1 and define $A = 0$ if $K \in V_0$ and $A = 1$ if $K \in V_1$. Let Y be the outcome and $Y(k)$ be the potential outcome for an individual if K had been k. Suppose that in an observational study for a set of covariates L we had $Y(k) \amalg K|L$ i.e. no confounding of the effect of K on Y conditional on covariates L. Suppose also the consistency assumption held such that $Y(k) = Y$ when $K = k$. An analyst who had dichotomized K so as to obtain a binary treatment A might then compute "causal effect" for dichotomized treatment A by calculating $\sum_l \mathrm{E}(Y|A = 1, l)pr(l) - \sum_l \mathrm{E}(Y|A = 0, l)pr(l)$. The following result re-expresses this quantities in terms of interventions on K. We state the result and then consider its interpretation.

Theorem 8. If $Y(k) \amalg K|L$ then

$$\sum_l \mathrm{E}(Y|A=1, l)pr(l) - \sum_l \mathrm{E}(Y|A=0, l)pr(l)$$
$$= \sum_l \mathrm{E}(Y(k)|l)pr(K=k|A=1, l)pr(l) - \sum_l \mathrm{E}(Y(k)|l)pr(K=k|A=0, l)pr(l).$$

This latter expression can itself be interpreted as a comparison in a randomized trial in which, within strata of covariates $L = l$, one arm is randomly assigned a "version of treatment" K from the observed distribution of K in the population amongst with $K \in V_1$ and $L = l$ (e.g. the distribution of minutes of exercise amongst those with $L = l$ who exercise at least 30 minutes) and the other arm is randomly assigned a "version of treatment" K from the observed distribution of K in the population amongst with $K \in V_0$ and $L = l$ (e.g. the distribution of minutes of exercise amongst those with $L = l$ who exercise less than 30 minutes). Note that nothing in the analysis above required that K itself be continuous; the variable K might indicate a complex set of treatment which are then dichotomized into a treatment variable A by partitioning the support of K into two sets, V_0 and V_1. A similar analysis is also applicable if K is categorized into some fixed number of categories, rather than dichotomized.

Note also that in some cases there may be ambiguity as to whether "treatment" precedes "version" as in the previous section or whether "version" precedes "treatment" as in this section. In the exercise example it is perhaps not unreasonable to argue that the number of minutes of exercise is constituted by a number of decisions (whether to exercise more than 5 minutes, whether to exercise more than 10 minutes, etc.). One of these decisions, namely whether to exercise more than 30 minutes, could be taken as A; once this is determined then there is still the question of which version of $A = 1$ (how many minutes above 30) or $A = 0$ (how many minutes below 30) is selected; this was how the issue was conceptualized in the previous section. In other cases, however, there is arguably less ambiguity. If the exposure $A = 1$ is experiencing high levels of loneliness and $A = 0$ experiencing low loneliness, then there are numerous decisions or interventions that may lead to high loneliness and it is difficult to conceive of these as following rather than preceding the high level of loneliness itself. In such cases the approach of this section will be of interest. In the next section we illustrate this approach with an empirical data analysis illustration.

6. Illustration

We illustrate some of the prior discussion with an example in which an exposure has been dichotomized. Loneliness (measured on the UCLA-R scale from 20 to 80) has been shown to prospectively predict depressive symptoms (measured on the CES-D scale from 0 to 60) even after control is made for baseline depressive symptoms and other covariates (Cacioppo et al., 2006). Longitudinal data available on loneliness and depressive symptoms in the Chicago Health, Aging, and Social Relations Study of 229 older adults; this data also include as covariates: age, gender, ethnicity, marital status, education, psychiatric conditions and psychiatric medications. Suppose a researcher were to median dichotomize measured at loneliness at follow-up 1 so as to define $A = 1$ when loneliness is greater than 35. A regression of depressive symptoms at follow-up 2 on dichotomized loneliness at follow-up 1, along with baseline loneliness, baseline depressive symptoms and baseline covariates gives an estimate of 2.44 (95% CI: $0.03, 4.85$) for dichotomized loneliness at follow-up 1. If we thought that baseline covariates (including baseline lonelines and depressive symptoms) were sufficient to control for confounding of the effect of loneliness at follow-up 1 on depressive symptoms at follow-up 2 then we could interpret this as an estimate of an intervention trial that, conditional on covariates, assigned each individual in one arm to a "version of treatment" of loneliness > 35 randomly drawn from the distribution of "versions of treatment" in the population of those with loneliness > 35 and assigned each individual in the other arm to a "version of treatment" of loneliness ≤ 35 randomly drawn from the distribution of "versions of treatment" in the population of those with loneliness ≤ 35. Note that the estimate only has this interpretation under the strong assumption of no unmeasured confounding, which may not be realistic here.

Note further that although under the assumption of no unmeasured confounding we can potentially interpret the effect of treatment in this manner, an intervention corresponding to the treatment effect we are supposedly estimating could not realistically be implemented in practice. Furthermore, whatever the underlying treatment variable K might be, to interpret the causal effect as the comparison from the randomized treatment regime trial described above we would have to control for all common causes of the underlying treatment K and the outcome. If we do not know what the underlying treatment K is, it is difficult to assess whether we have indeed controlled for all relevant confounders. We have discussed in greater detail these points and their implications for epidemiologic research elsewhere (Hernán and VanderWeele, 2010).

7. Discussion

In this article we have described how the potential outcomes framework can be extended to allow for multiple versions of treatment, which are present to varying degrees in both observational studies and in randomized clinical trials. Multiple versions of treatment are arguably present to varying degrees in both observational studies and in randomized clinical trials. In clinical trials, guidelines are often given to reduce the number or relevance of versions but it is generally not possible to eliminate this problem entirely (Hernán and VanderWeele, 2011).Fortunately, as we have seen, even with multiple versions of treatment, it is possible to use the ordinary estimators for causal effects to compare the effects of treatments on average under no-unmeasured-confounding assumptions. Ordinary estimators can be interpreted as the causal effects of well-defined interventions that mimic the assignment of versions of treatment in the study population. For such an interpretation in an observational study control must, however, in general be made for common causes of treatment and version; in a randomized trial there will be no such common causes. Although the ordinary estimators have an interpretation as an overall causal effect, multiple versions of treatment still renders ambiguous statements such as “treatment is on average better than control” since these statements will always be with reference to the current policies for assigning versions. Even if treatment is better on average than control under current policies for assigning versions, it is nevertheless possible that certain versions of control, if administered to an entire population, would be better than administering certain versions (or even all versions) of treatment. This could arise if the most effective version of the control were generally infrequently assigned. Analysis of the effects of a specific version of treatment (Theorem 1) or of varying regimes and policies (as considered in section 4) can be useful in assessing this possibility.

The contributions in this paper have attempted to extend the potential outcomes framework to allow for multiple versions of treatment. Recent work in causal inference has also attempted to extend the standard potential outcomes notation to accommodate possible interference between units (Sobel, 2006; Hong and Raudenbush, 2006; Rosenbaum, 2007; Hudgens and Halloran, 2008; VanderWeele, 2010; Tchetgen Tchetgen and VanderWeele, 2010). We would like to conclude this paper by drawing some parallels between the existing work on interference and our discussion above concerning multiple versions of treatment. First, both the no-interference assumption and the no-multiple-versions-of-treatment assumption are concealed by the notation $Y_j(a)$; these two assumptions are often not stated explicitly but are implicitly assumed to hold when using potential outcomes notation such as $Y_j(a)$; such notation is in general only justified under the assumptions of no-interference and no-multiple-versions-of-treatment. Second, with both the no-interference assumption and the no-multiple-versions-of-treatment assumption, although the traditional potential outcomes framework presupposes these assumptions, the framework and notation can in fact be extended so as to allow for potential violations; in the case of interference, the potential outcomes notation can be extended so as to allow the potential outcome of one individual to depend on the treatments received by other individuals; in the case of multiple versions of treatment, the notation can be expanded so that an individual may have different potential outcomes for each possible version of treatment. Third, in certain settings, violations of the no-interference assumption or the no-multiple-versions-of-treatment assumption can be ignored; Rosenbaum (2007) showed that the no-interference assumption could be ignored in certain randomized experiments; in our discussion above we have seen that if covariates are available to adjust not just for treatment-outcome confounding but also for "treatment-version confounding" then the multiple versions of treatment can be ignored in the estimation of average causal effects (it is not necessary to have data on which individuals received which version). Fourth, with both the no-interference assumption and the no-multiple-versions-of-treatment assumption, once notation has been introduced to expand the potential outcomes framework in order to accommodate violations, then this new notation can give rise to new questions of theoretical and substantive interest; notation accommodating interference gives rise to questions of the identification and estimation of spillover effects; notation accommodating multiple versions of treatment gives rise to questions about hypothetical interventions on the version of treatment to address policy relevant questions about resource allocation and assignment.

The extension of the potential outcomes framework to address interference and spillover effects has been of use in a variety of substantive contexts (Hong and Raudenbush, 2006; Sobel, 2006; Hudgens and Halloran, 2008). We hope that the contributions in this article will similarly clarify and extend the possibilities for causal inference when multiple versions of treatment are present.

References.

Acute Respiratory Distress Syndrome Network. (2000). Ventilation with lower tidal volumes as compared with traditional tidal volumes for acute lung injury and the acute respiratory distress syndrome. *New England Journal of Medicine* **342**, 1301-8.

Cacioppo, J., Hughes, M., Waite, L., Hawkley, L., & Thisted, R. (2006). Loneliness as a specific risk-factor for depressive symptoms: Cross-sectional and longitudinal analyses. *Psychology and Aging* **21**, 140-151.

Cole, S. R. and Frangakis, C. E. (2009). The consistency assumption in causal inference: a definition or an assumption? *Epidemiology* **20**, 3-5.

Cox, D. R., (1958). *Planning of Experiments*. New York: John Wiley & Sons.

Geneletti, S. (2007). Identifying direct and indirect effects in a non-counterfactual framework. *Journal of the Royal Statistical Soceity, Series B*, 69:199-216.

Hernán, M.A. and VanderWeele, T.J. (2010). Compound treatments and transportability of causal inference. *Epidemiology*, in press.

Hong, G. and Raudenbush, S. W. (2006). Evaluating kindergarten retention policy: A case study of causal inference for multilevel observational data. *Journal of the American Statistical Association* **101**, 901-910.

Hudgens, M. G. and Halloran, M. E. (2008). Towards causal inference with interference. *Journal of the American Statistical Association* **103**, 832-842.

Neyman, J. (1923). Sur les applications de la thar des probabilities aux experiences Agaricales: Essay des principle. Excerpts reprinted (1990) in English (D. Dabrowska and T. Speed, Trans.) in *Statistical Science* **5**, 463-472.

Neyman, J. (1935). Statistical problems in agricultural experimentation. *Journal of the Royal Statistical Society, II* **2**, 107-154.

Pearl, J. (2001). Direct and indirect effects. In *Proceedings of the Seventeenth Conference on Uncertainty and Artificial Intelligence*. San Francisco: Morgan Kaufmann, 411-420.

Pearl, J. (2009). *Causality: Models, Reasoning, and Inference*. Cambridge: Cambridge University Press.

Pearl, J. and Robins, J. M. (1995). Probabilistic evaluation of sequential plans from causal models with hidden variables. In *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, 444-453.

Robins, J. M. (1986). A new approach to causal inference in mortality studies with sustained exposure period - application to control of the healthy worker survivor effect. *Mathematical Modelling* **7**,1393-1512.

Robins, J.M. (2003). Semantics of causal DAG models and the identification of direct and indirect effects. In Highly Structured Stochastic Systems, Eds. P. Green, N.L. Hjort, and S. Richardson, 70-81. Oxford University Press, New York.

Robins, J.M. & Greenland, S. (2000). Comment on: "Causal inference without counterfactuals" by A.P. Dawid. *Journal of the American Statistical Association* **95**, 477-82.

Rosenbaum, P. R. (2007). Interference between units in randomized experiments. *Journal of the American Statistical Association* **102**, 191-200.

Rosenbaum, P. R. and Rubin, D. B. (1983). Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome. *Journal of the Royal Statistical Society, Series B* **45**, 212-218.

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

Rubin, D. B. (1980). Comment on: "Randomization analysis of experimental data in the fisher randomization test" by D. Basu. *Journal of the American Statistical Association* **75**, 591-593.

Rubin, D. B. (1986). Which ifs have causal answers? Comment on: "Statistics and causal inference" by P. Holland. *Journal of the American Statistical Association* **81**, 961-962.

Rubin, D. B. (1990). Formal modes of statistical inference for causal effects. *Journal of Statistical Planning and Inference* **25**, 279-292.

Schafer, J. L. and Kang, J. (2009). Causal modeling when the treatment is a latent class. Abstract. *Joint Statistical Meetings*, 2009.

Sobel, M. E. (2006) What do randomized studies of housing mobility demonstrate?: Causal inference in the face of interference. *Journal of the American Statistical Association* **101**, 1398-1407.

Taubman, S. L., Robins, J. M., Mittleman, M. A. and Hernán, M, A. (2008). Alternative approaches to estimating the effects of hypothetical interventions. In: *Proceedings of the 2008 Joint Statistical Meetings*; Alexandria, VA.

Tchetgen Tchetgen, E.J. and VanderWeele, T.J. (2010). On causal inference in the presence of interference. *Statistical Methods in Medical Research – Special Issue on Causal Inference*, in press. Published online Nov. 10, 2010, doi: 10.1177/0962280210386779.

VanderWeele, T. J. (2009). Further remarks concerning the consistency assumption. *Epidemiology* **20**, 880-883.

VanderWeele, T.J. (2010). Direct and indirect effects for neighborhood-based clustered and longitudinal data. *Sociological Methods and Research*, **38**, 515-544.

Supplementary Materials for "Causal inference under multiple versions of treatment" by TJ VanderWeele and MA Hernán.

Proofs of Theorems 1-8.

Proof of Theorem 1. For any $a \in \mathcal{A}, k^a \in \mathcal{K}^a$ we have that

$$\begin{aligned} \mathrm{E}\{Y(a,k^a)\} &= \sum\nolimits_c \mathrm{E}\{Y(a,k^a)|C=c\}pr(c) \\ &= \sum\nolimits_c \mathrm{E}\{Y(a,k^a)|A=a,K^a=k^a,C=c\}pr(c) \text{ by (8)} \\ &= \sum\nolimits_c \mathrm{E}\{Y|A=a,K^a=k^a,C=c\}pr(c) \text{ by consistency.}\blacksquare \end{aligned}$$

Proof of Theorem 2. We have that

$$
\begin{aligned}
\mathrm{E}\{Y(a,K^a)|A=a\} &= \sum\nolimits_{k^a} \mathrm{E}\{Y(a,K^a)|A=a,K^a=k^a\}P(K^a=k^a|A=a) \\
&= \sum\nolimits_{k^a} \mathrm{E}(Y|A=a,K^a=k^a)P(K^a=k^a|A=a) \text{ by consistency} \\
&= \mathrm{E}(Y|A=a).
\end{aligned}
$$

Note that if there is only one version of treatment for the control condition, $A=0$, then $\mathcal{K}^0 = \{1\}$ and the only potential outcome for each individual under the control condition is $Y_j(0) \equiv Y_j(0,K_j^0) = Y_j(0,k^0=1)$. We have that

$$
\begin{aligned}
\mathrm{E}\{Y(0)|A=a\} &= \mathrm{E}\{Y(0,k^0=1)|A=a\} \\
&= \sum\nolimits_c \mathrm{E}\{Y(0,k^0=1)|A=a\}pr(c|A=a) \\
&= \sum\nolimits_c \mathrm{E}\{Y(0,k^0=1)|A=0,K^0=1,C=c\}pr(c|A=a) \text{ since } Y(0) \amalg \{A,K^0\}|C \\
&= \sum\nolimits_c \mathrm{E}\{Y|A=0,k^0=1,C=c\}pr(c|A=a) \text{ by consistency} \\
&= \sum\nolimits_c \mathrm{E}\{Y|A=0,C=c\}pr(c|A=a)
\end{aligned}
$$

where the final equality holds because when $A=0$ we have that $K^0=1$ since there is only one version of treatment.■

Proof of Theorem 3. For any $a \in \mathcal{A}$, we have that
$\mathrm{E}\{Y(a,K^a(a))\} = \mathrm{E}\{Y(a)\}$

$$
\begin{aligned}
&= \sum\nolimits_{c,w} \mathrm{E}\{Y(a)|c,w\}pr(c,w) \\
&= \sum\nolimits_{c,w} \mathrm{E}\{Y(a)|a,c,w\}pr(c,w) \text{ by (9)} \\
&= \sum\nolimits_{c,w} \mathrm{E}\{Y(a,K^a(a))|a,c,w\}pr(c,w) \\
&= \sum\nolimits_{c,w,k^a} \mathrm{E}\{Y(a,k^a)|a,K^a(a)=k^a,c,w\}pr\{K^a(a)=k^a|a,c,w\}pr(c,w) \\
&= \sum\nolimits_{c,w,k^a} \mathrm{E}\{Y(a,k^a)|a,K^a=k^a,c,w\}pr(K^a=k^a|a,c,w)pr(c,w) \text{ by consistency for } K \\
&= \sum\nolimits_{c,w,k^a} \mathrm{E}\{Y|a,K^a=k^a,c,w\}pr(K^a=k^a|a,c,w)pr(c,w) \text{ by consistency for } Y \\
&= \sum\nolimits_{c,w} \mathrm{E}\{Y|a,c,w\}pr(c,w).
\end{aligned}
$$

This completes the proof.■

Proof of Theorem 4. We have that

$$\begin{aligned}
\mathrm{E}\{Y(a,k^a)|A &= a^*\} = \sum\nolimits_c \mathrm{E}\{Y(a,k^a)|A=a^*, C=c\}pr(c|a^*) \\
&= \sum\nolimits_c \mathrm{E}\{Y(a,k^a)|A=a, K^a=k^a, C=c\}pr(c|a^*) \text{ by (8)} \\
&= \sum\nolimits_c \mathrm{E}\{Y|A=a, K^a=k^a, C=c\}pr(c|a^*) \text{ by consistency.}\blacksquare
\end{aligned}$$

Proof of Theorem 5. We have that
$\mathrm{E}\{Y(a,G^a)|A=a^*\}$

$$\begin{aligned}
&= \sum\nolimits_c \mathrm{E}\{Y(a,G^a)|A=a^*, C=c\}pr(c|a^*) \\
&= \sum\nolimits_{c,k^a} \mathrm{E}\{Y(a,k^a)|G^a=k^a, A=a^*, C=c\}pr(G^a=k^a|A=a^*, C=c)pr(c|a^*) \\
&= \sum\nolimits_{c,k^a} \mathrm{E}\{Y(a,k^a)|A=a^*, C=c\}pr(K^a=k^a|A=a, C=c)pr(c|a^*) \\
&= \sum\nolimits_{c,k^a} \mathrm{E}\{Y(a,k^a)|A=a, K^a=k^a, C=c\}pr(K^a=k^a|A=a, C=c)pr(c|a^*) \text{ by (8)} \\
&= \sum\nolimits_{c,k^a} \mathrm{E}\{Y|A=a, K^a=k^a, C=c\}pr(K^a=k^a|A=a, C=c)pr(c|a^*) \text{ by consistency.}\blacksquare
\end{aligned}$$

Proof of Theorem 6. We have that
$\mathrm{E}\{Y(a,G_j^{a|a^*})|A=a^*\}$

$$\begin{aligned}
&= \sum\nolimits_{c,w} \mathrm{E}\{Y(a,G_j^{a|a^*})|A=a^*, C=c, W=w\}pr(c,w|a^*) \\
&= \sum\nolimits_{c,w,k^a} \mathrm{E}\{Y(a,k^a)|G_j^{a|a^*}=k^a, a^*, c, w\}pr(G_j^{a|a^*}=k^a|a^*, c, w)pr(c,w|a^*) \\
&= \sum\nolimits_{c,w,k^a} \mathrm{E}\{Y(a,k^a)|a^*, c, w\}pr(K^a(a)=k^a|a^*, c, w)pr(c,w|a^*) \\
&= \sum\nolimits_{c,w,k^a} \mathrm{E}\{Y(a,k^a)|a^*, c, w\}pr(K^a(a)=k^a|a, c, w)pr(c,w|a^*) \text{ by (10)} \\
&= \sum\nolimits_{c,w,k^a} \mathrm{E}\{Y(a,k^a)|a, k^a, c, w\}pr(K^a(a)=k^a|a, c, w)pr(c,w|a^*) \text{ by (8)} \\
&= \sum\nolimits_{c,w,k^a} \mathrm{E}\{Y|a, k^a, c, w\}pr(K^a=k^a|a, c, w)pr(c,w|a^*) \text{ by consistency.}\blacksquare
\end{aligned}$$

Proof of Theorem 7. We have that

$$\begin{aligned}
\mathrm{E}\{Y(a, G^{a,s})|A &= a, S = s\} = \sum\nolimits_{c} \mathrm{E}\{Y(a, G^{a,s})|A = a, C = c, S = s\}pr(c|a, s) \\
&= \sum\nolimits_{c,k^a} \mathrm{E}\{Y(a, k^a)|G^{a,s} = k^a, a, c, s\}pr(G^{a,s} = k^a|a, c, s')pr(c|a, s) \\
&= \sum\nolimits_{c,k^a} \mathrm{E}\{Y(a, k^a)|a, c, s\}pr(K^a = k^a|a, c, s')pr(c|a, s) \\
&= \sum\nolimits_{c,k^a} \mathrm{E}\{Y(a, k^a)|a, k^a, c, s\}pr(K^a = k^a|a, c, s')pr(c|a, s) \text{ by (8)} \\
&= \sum\nolimits_{c,k^a} \mathrm{E}(Y|a, k^a, c, s)pr(K^a = k^a|a, c, s')pr(c|a, s) \text{ by consistency.} \blacksquare
\end{aligned}$$

Proof of Theorem 8
If $Y(k) \amalg K|L$ then

$$\begin{aligned}
& \sum\nolimits_{l} \mathrm{E}(Y|A = 1, l)pr(l) - \sum\nolimits_{l} \mathrm{E}(Y|A = 0, l)pr(l) \\
= & \sum\nolimits_{l} \mathrm{E}(Y|A = 1, K = k, l)pr(K = k|A = 1, l)pr(l) - \sum\nolimits_{l} \mathrm{E}(Y|A = 0, K = k, l)pr(K = k|A = 0, l)pr(l \\
= & \sum\nolimits_{l} \mathrm{E}(Y|K = k, l)pr(K = k|A = 1, l)pr(l) - \sum\nolimits_{l} \mathrm{E}(Y|K = k, l)pr(K = k|A = 0, l)pr(l) \\
= & \sum\nolimits_{l} \mathrm{E}(Y(k)|K = k, l)pr(K = k|A = 1, l)pr(l) - \sum\nolimits_{l} \mathrm{E}(Y(k)|K = k, l)pr(K = k|A = 0, l)pr(l) \\
= & \sum\nolimits_{l} \mathrm{E}(Y(k)|l)pr(K = k|A = 1, l)pr(l) - \sum\nolimits_{l} \mathrm{E}(Y(k)|l)pr(K = k|A = 0, l)pr(l)
\end{aligned}$$

where the first equality follows by iterated expectations, the second because K contains all the information in A, the third from consistency and the fourth because $Y(k) \amalg K|L$.$\blacksquare$

Example of Treatment-Version Confounding.

Let $A = 1$ denote surgery and $A = 0$ denote the control (no surgery). Suppose there is only one version of $A = 0$ (no surgery) but two versions of surgery: surgeon 1 ($k^1 = 1$) and surgeon 2 ($k^1 = 2$). Suppose that there are no confounders C that affect both the outcome Y and either treatment or version but that there is a binary treatment-version confounder W with $W = 0$ and $W = 1$ indicating two different health plans. Suppose $P(W = 1) = 0.5$; $P(A = 1|W = 0) = 0.2$; $P(A = 1|W = 1) = 0.6$ so that by Bayes' Theorem, $P(W = 1|A = 1) = 3/4$ and $P(W = 1|A = 0) = 1/3$. Suppose also $P(K^1 = 1|A = 1, W = w) = (420 + 300W)/1000$. Finally, suppose $P(Y = 1|A = 0, W = w) = 1/2$ and $P(Y = 1|A = 1, K^1 = k^1, W = w) = 1/2 + k^1/5$. Note that W affects both A and K^1 but W has no effect on Y except through A and K^1.

Now if the entire population were given surgery ($A = 1$) then the proportion of the population with $K^1 = 1$ would be $\{420 + 300\,\mathrm{E}(W)\}/1000 = 0.57$ and the proportion with $K^1 = 2$ would thus be 0.43. The proportion with $Y = 1$ would be $1/2 + \mathrm{E}(K^1)/5 = 1/2 + (0.57 + 2 * 0.43)/5 = 0.786$. If the entire population were not given surgery ($A = 0$) the proportion with $Y = 1$ would be 0.5. The true overall average causal effect of surgery in this population is thus $0.786 - 0.5 = 0.286$.

Suppose now that no information is available on version of treatment. Suppose that we did not control for W and simply computed $\mathrm{E}(Y|A=1) - \mathrm{E}(Y|A=0)$. We would obtain:

$$
\begin{aligned}
& \mathrm{E}(Y|A=1) - \mathrm{E}(Y|A=0) \\
= \; & \sum\nolimits_w \mathrm{E}(Y|A=1,w)pr(w|A=1) - \sum\nolimits_w \mathrm{E}(Y|A=0,w)pr(w|A=0) \\
= \; & \sum\nolimits_w \sum\nolimits_{k^1} \mathrm{E}(Y|A=1,k^1,w)pr(k^1|A=1,W=w)pr(w|A=1) - \sum\nolimits_w \mathrm{E}(Y|A=0,w)pr(w|A=0) \\
= \; & \sum\nolimits_w \sum\nolimits_{k^1} (1/2 + k^1/5)pr(k^1|A=1,W=w)pr(w|A=1) - 1/2 \\
= \; & \sum\nolimits_w \sum\nolimits_{k^1} (k^1/5)pr(k^1|A=1,W=w)pr(w|A=1) \\
= \; & (1/5)\sum\nolimits_w \mathrm{E}(K^1|A=1,W=w)pr(w|A=1) \\
= \; & (1/5)\sum\nolimits_w [(420+300W)/1000 + 2*\{1-(420+300W)/1000\}]pr(w|A=1) \\
= \; & (1/5)\sum\nolimits_w \{2-(420+300W)/1000\}pr(w|A=1) \\
= \; & (1/5)[2-\{420+300\,\mathrm{E}(W|A=1)\}/1000] \\
= \; & (1/5)[2-\{420+300*3/4\}/1000] = 0.271.
\end{aligned}
$$

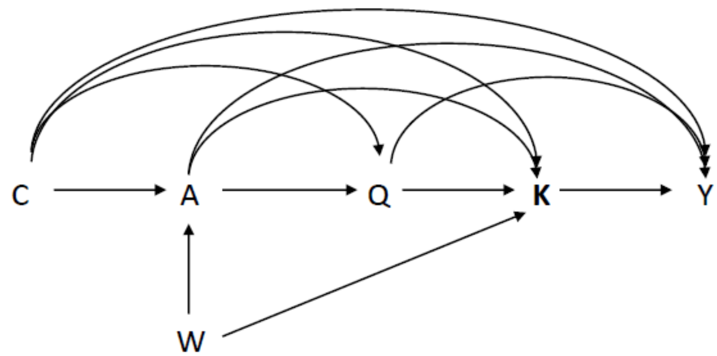
We would get a biased estimate of the overall average causal effect of surgery. If we did control for W and computed $\sum_w \{\mathrm{E}(Y|A=1,w) - \mathrm{E}(Y|A=0,w)\}pr(w)$ we would obtain:

$$
\begin{aligned}
& \sum\nolimits_w \mathrm{E}(Y|A=1,w)pr(w) - \mathrm{E}(Y|A=0,w)pr(w) \\
= \; & \sum\nolimits_w \sum\nolimits_{k^1} \mathrm{E}(Y|A=1,k^1,w)pr(k^1|A=1,W=w)pr(w) - \sum\nolimits_w \mathrm{E}(Y|A=0,w)pr(w) \\
= \; & \sum\nolimits_w \sum\nolimits_{k^1} (1/2 + k^1/5)pr(k^1|A=1,W=w)pr(w) - 1/2 \\
= \; & \sum\nolimits_w \sum\nolimits_{k^1} (k^1/5)pr(k^1|A=1,W=w)pr(w) \\
= \; & (1/5)\sum\nolimits_w \mathrm{E}(K^1|A=1,W=w)pr(w) \\
= \; & (1/5)\sum\nolimits_w [(420+300W)/1000 + 2*\{1-(420+300W)/1000\}]pr(w) \\
= \; & (1/5)\sum\nolimits_w \{2-(420+300W)/1000\}pr(w) \\
= \; & (1/5)[2-\{420+300\,\mathrm{E}(W)\}/1000] \\
= \; & (1/5)[2-\{420+300*1/2\}/1000] = 0.286.
\end{aligned}
$$

We get a correct estimate of the overall average causal effect of surgery is we control for the treatment-version confounder W but a biased estimate if we do not control for it.

Analogous Results Under Time-Dependent Confounding.

Suppose now that there is an effect, Q, of treatment A that affects both version $\mathbf{K}$ and the outcome Y as in the Appendix Figure. We will continue to let W denote a set of variables that affects only treatment A and version $\mathbf{K}$.



Appendix Figure. Time-dependent confounding in which an effect, Q, of treatment A, may affect both version K and outcome Y.

We replace assumption (8) with

$$Y(a, k^a) \perp\!\!\!\perp A|C \text{ for all } a \in \mathcal{A}, k^a \in \mathcal{K}^a \tag{A1}$$
$$Y(a, k^a) \perp\!\!\!\perp \mathbf{K}|(C, A, Q) \text{ for all } a \in \mathcal{A}, k^a \in \mathcal{K}^a \tag{A2}$$

Under these two assumptions, the effect of the version of treatment remains identified but data must be available on both version K^a and on the time-dependent confounder Q as stated in the following result which provides the analogue to Theorem 1 under time-dependent confounding. The proof is somewhat analogous to that for "controlled direct effects" in the context of mediation with a time-dependent confounder.

Theorem 9. Under assumptions (A1) and (A2),

$$\mathrm{E}\{Y(a, k^a)\} = \sum_{c,q} \mathrm{E}\{Y|A = a, K^a = k^a, c, q\} pr(q|A = a, c) pr(c)$$

Proof. For any $a \in \mathcal{A}, k^a \in \mathcal{K}^a$ we have that

$$\begin{aligned}
\mathrm{E}\{Y(a, k^a)\} &= \sum_c \mathrm{E}\{Y(a, k^a)|c\} pr(c) \\
&= \sum_c \mathrm{E}\{Y(a, k^a)|A = a, c\} pr(c) \text{ by (A1)} \\
&= \sum_{c,q} \mathrm{E}\{Y(a, k^a)|A = a, c, q\} pr(q|A = a, c) pr(c) \\
&= \sum_{c,q} \mathrm{E}\{Y(a, k^a)|A = a, K^a = k^a, c, q\} pr(q|A = a, c) pr(c) \text{ by (A2)} \\
&= \sum_{c,q} \mathrm{E}\{Y|A = a, K^a = k^a, c, q\} pr(q|A = a, c) pr(c) \text{ by consistency.} \blacksquare
\end{aligned}$$

Even in the presence of time-dependent confounding as in the Appendix Figure, assumption (9) in the text that $Y(a) \amalg A|\{C, W\}$ for all a will still hold and the overall causal effect will be identified by the proof of Theorem 3 given above. As before, data on version of treatment is thus not necessary to estimate overall treatment effects. Throughout the paper and in the Appendix we have, however, assumed point treatment. If treatment is time-varying then the version of treatment may serve as a confounder for the effect of subsequent treatment and data would then be needed on the version of treatment for the purposes of confounding control. The development of a formal analytic framework for this setting is left to future research.

In the text, Theorems 5 and 6 considered the effects on those with $A = a^*$ of intervening to set $A = a$ with version randomly set to the distribution of those with $A = a$ (Theorem 5) or to the distribution of those with $A = a^*$ had they been given treatment $A = a$. Theorems 10 and 11 give analogous results under time-dependent confounding. Theorem 10 requires assumptions (A1) and (A2). Theorem requires assumptions (A1) and (A2) along with assumption (10) in the text that $K^a(a) \amalg A|\{C, W\}$ for all a.

Theorem 10. For individuals j with $A_j = a^* \neq a$, let G^a_j be a random variable with distribution defined by $pr(K^a = k^a|A = a, C = C_j)$. If for all j such that $A_j = a^*$, the potential outcome $Y_j(a, k^a)$ is well defined for all $k^a \in supp(G^a_j)$ and assumptions (A1) and (A2) hold then

$$\begin{aligned} & \mathrm{E}\{Y(a, G^a)|A = a^*\} \\ = & \textstyle\sum_{c,k^a,q} \mathrm{E}\{Y|A = a, K^a = k^a, c, q\} pr(q|A = a, c) pr(K^a = k^a|A = a, c) pr(c|a^*). \end{aligned}$$

Proof. We have that
$\mathrm{E}\{Y(a, G^a)|A = a^*\}$

$$\begin{aligned}
= & \textstyle\sum_{c} \mathrm{E}\{Y(a, G^a)|A = a^*, c\} pr(c|a^*) \\
= & \textstyle\sum_{c,k^a} \mathrm{E}\{Y(a, k^a)|A = a^*, c\} pr(G^a = k^a|A = a^*, c) pr(c|a^*) \\
= & \textstyle\sum_{c,k^a} \mathrm{E}\{Y(a, k^a)|A = a, c\} pr(K^a = k^a|A = a, c) pr(c|a^*) \text{ by (A1)} \\
= & \textstyle\sum_{c,k^a,q} \mathrm{E}\{Y(a, k^a)|A = a, c, q\} pr(q|A = a, c) pr(K^a = k^a|A = a, c) pr(c|a^*) \\
= & \textstyle\sum_{c,k^a,q} \mathrm{E}\{Y(a, k^a)|A = a, K^a = k^a, c, q\} pr(q|A = a, c) pr(K^a = k^a|A = a, c) pr(c|a^*) \text{ by (A2)} \\
= & \textstyle\sum_{c,k^a,q} \mathrm{E}\{Y|A = a, K^a = k^a, c, q\} pr(q|A = a, c) pr(K^a = k^a|A = a, c) pr(c|a^*) \text{ by consistency.} \blacksquare
\end{aligned}$$

Theorem 11. If for all individuals j with $A_j = a^* \neq a$, the potential outcome $K^a_j(a)$ is well defined then let $G^{a|a^*}_j$ be a random variable with distribution defined by $pr(K^a(a) = k^a|A = a^*, C = C_j, W = W_j)$. If for all j such that $A_j = a^*$, the potential outcome $Y_j(a, k^a)$ is well defined for all $k^a \in supp(G_j^{a|a^*})$ and if assumptions (A1), (A2) and (10) hold then

$$\begin{aligned} & \mathrm{E}\{Y(a, G_j^{a|a^*})|A = a^*\} \\ = & \sum\nolimits_{c,w,k^a,q} \mathrm{E}\{Y|a, k^a, c, w, q\} pr(q|a, c, w) pr(K^a(a) = k^a|a, c, w) pr(c, w|a^*). \end{aligned}$$

Proof. We have that
$\mathrm{E}\{Y(a, G_j^{a|a^*})|A = a^*\}$

$$\begin{aligned} = & \sum\nolimits_{c,w} \mathrm{E}\{Y(a, G_j^{a|a^*})|A = a^*, c, w\} pr(c, w|a^*) \\ = & \sum\nolimits_{c,w,k^a} \mathrm{E}\{Y(a, k^a)|a^*, c, w\} pr(G_j^{a|a^*} = k^a|a^*, c, w) pr(c, w|a^*) \\ = & \sum\nolimits_{c,w,k^a} \mathrm{E}\{Y(a, k^a)|a, c, w\} pr(K^a(a) = k^a|a^*, c, w) pr(c, w|a^*) \text{ by (A1)} \\ = & \sum\nolimits_{c,w,k^a} \mathrm{E}\{Y(a, k^a)|a, c, w\} pr(K^a(a) = k^a|a, c, w) pr(c, w|a^*) \text{ by (10)} \\ = & \sum\nolimits_{c,w,k^a,q} \mathrm{E}\{Y(a, k^a)|a, c, w, q\} pr(q|a, c, w) pr(K^a(a) = k^a|a, c, w) pr(c, w|a^*) \\ = & \sum\nolimits_{c,w,k^a,q} \mathrm{E}\{Y(a, k^a)|a, k^a, c, w, q\} pr(q|a, c, w) pr(K^a(a) = k^a|a, c, w) pr(c, w|a^*) \text{ by 8} \\ = & \sum\nolimits_{c,w,k^a,q} \mathrm{E}\{Y|a, k^a, c, w, q\} pr(q|a, c, w) pr(K^a(a) = k^a|a, c, w) pr(c, w|a^*) \text{ by consistency.} \blacksquare \end{aligned}$$