

# Compound Treatments and Transportability of Causal Inference

Miguel A. Hernán<sup>a,b</sup> and Tyler J. VanderWeele<sup>a,c</sup>

**Abstract:** Ill-defined causal questions present serious problems for observational studies—problems that are largely unappreciated. This paper extends the usual counterfactual framework to consider causal questions about compound treatments for which there are many possible implementations (for example, “prevention of obesity”). We describe the causal effect of compound treatments and their identifiability conditions, with a special emphasis on the consistency condition. We then discuss the challenges of using the estimated effect of a compound treatment in one study population to inform decisions in the same population and in other populations. These challenges arise because the causal effect of compound treatments depends on the distribution of the versions of treatment in the population. Such causal effects can be unpredictable when the versions of treatment are unknown. We discuss how such issues of “transportability” are related to the consistency condition in causal inference. With more carefully framed questions, the results of epidemiologic studies can be of greater value to decision-makers.

(*Epidemiology* 2011;22: 368–377)

Criticisms of observational analyses are usually focused on confounding and other biases, and less often on the adequacy of the questions asked by those analyses. Yet all of us conducting observational studies occasionally report answers to questions that are, if closely examined, of questionable interest for decision-making. An example of such questions is: “How many excess deaths are attributable to obesity each year in my country?”

The relevance of questions for decision-making is intimately linked to the definition of treatment, the existence of multiple versions of treatment, and the consistency of the potential or counterfactual outcomes. Such consistency has traditionally been a central issue in philosophical discussions

about causality.<sup>1</sup> To ensure relevant causal questions, some authors have proposed restricting causal inferences to treatments for which a hypothetical intervention can be unambiguously specified. This idea is encapsulated in the adage “no causation without manipulation.”<sup>2</sup> More recently several authors—statisticians, epidemiologists, and computer scientists—have continued the discussion on the implications of ill-defined causal questions.<sup>3–12</sup>

This paper proposes a framework for causal questions about treatments with multiple versions, which we refer to as compound treatments. We adopt the perspective of a decision-maker (eg, physician, patient, public health officer, policy-maker) who needs to make a decision based on observational data, rather than a philosopher interested in causality, or a scientist interested in explaining the world.<sup>11</sup> We start by defining compound treatments with multiple versions in both randomized experiments and observational studies. We then review identification of causal effects under multiple versions of treatment, with a special focus on the consistency condition. We describe the challenges that arise when trying to use the effect of a compound treatment estimated in a population to make decisions in another population, both when the versions of treatment are known and when they are unknown. We conclude with a discussion of how such issues of “transportability” are related to the consistency condition in causal inference. For simplicity, we assume no measurement error.

## COMPOUND TREATMENTS IN RANDOMIZED EXPERIMENTS

Consider a randomized experiment to estimate the causal effect of a dichotomous non-time-varying treatment  $A$  on the 5-year risk of death  $Y$  (1: dead, 0: alive). All individuals adhered to the assigned treatment throughout the entire follow-up. Let  $Y_i^a$  be individual  $i$ 's potential or counterfactual outcome under treatment value  $a$ . The average causal effect of treatment  $A$  on outcome  $Y$  is defined as  $\beta_A = E[Y^a] - E[Y^{a'}]$ , where  $a$  and  $a'$  are the 2 values that treatment  $A$  can take. For example, suppose that  $A$  represents aspirin dose and that individuals are assigned to either a daily tablet of 150 mg of aspirin plus some particular combination of inactive ingredients ( $A = 1$ ), or a tablet that is identical except it contains no aspirin ( $A = 0$ ). Then the average causal effect of aspirin that compares  $A = 1$  with  $A = 0$  is  $\beta_A = E[Y^{a=1}] - E[Y^{a=0}]$ . As

Submitted May 26, 2010; accepted 26 October 2010; posted 11 March 2011.  
From the <sup>a</sup>Department of Epidemiology, Harvard School of Public Health, Boston, MA; <sup>b</sup>Harvard-MIT Division of Health Sciences and Technology, Boston, MA; and <sup>c</sup>Department of Biostatistics, Harvard School of Public Health, Boston, MA.

Supported by NIH grants HL080644, AI073127, and HD060696.

**Editors' note:** A commentary on this article appears on page 378.

Correspondence: Miguel Hernán, Department of Epidemiology, Harvard School of Public Health, 677 Huntington Avenue, Boston, MA 02115.  
E-mail: miguel\_hernan@post.harvard.edu.

Copyright © 2011 by Lippincott Williams & Wilkins

ISSN: 1044-3983/11/2203-0368

DOI: 10.1097/EDE.0b013e3182109296

another example, suppose that treatment  $A$  represents daily duration of exercise, and that individuals are assigned to exercise either 60 minutes or 10 minutes per day. Then the average causal effect of exercise that compares  $A = 60$  with  $A = 10$  is  $\beta_A = E[Y^{a=60}] - E[Y^{a=10}]$ .

The causal graph in Figure 1 represents a randomized experiment with treatment  $A$ . The graph does not include any common causes of the treatment  $A$  and the outcome  $Y$  because exchangeability is expected when treatment is randomized. Therefore,  $E[Y^a] = E[Y|A = a]$ , and the average causal effect  $\beta_A = E[Y^a] - E[Y^{a'}]$  is unbiasedly estimated without any adjustment for covariates.

Now suppose investigators are interested in estimating the effect of treatment  $R$ , and hence they randomly assigned individuals to either  $R = 1$ : “exercise at least 30 minutes daily,” or  $R = 0$ : “exercise less than 30 minutes daily.” Individuals are free to choose the actual duration of exercise as long as it is consistent with their assignment. Thus, the duration of daily exercise will take a range of values from 30 to, say, 180 minutes in group  $R = 1$ , and from 0 to 29 minutes in group  $R = 0$ . Each of the possible durations 30, 31, 32... minutes can be viewed as a different version of the treatment  $R = 1$ , and similarly each of the possible durations 0, 1, 2, ... 29 minutes can be viewed as a different version of the treatment  $R = 0$ . More formally, let  $A_i(r)$  be the version of treatment  $R = r$  that individual  $i$  received.  $A(r)$  takes values in some set  $\mathcal{A}(r) = \{[1, \dots, n(r)]\}$ . In our example,  $\mathcal{A}(1)$  is the set  $\{30, 31, 32, \dots\}$  (with appropriate offset by starting at 30 rather than 1) indicating all possible durations of exercise greater than or equal to 30 minutes, and  $\mathcal{A}(0)$  is the set  $\{0, 1, 2, \dots, 29\}$  including all durations of less than 30 minutes. We refer to  $R$  as a compound treatment because multiple values  $a(r)$  can be mapped onto a single value  $R = r$ . Another example of compound treatment arises in randomized experiments with imperfect adherence to treatment: the intention-to-treat effect of treatment assignment  $R$  depends on the patterns of adherence to  $R$  (the versions of treatment).

The causal graph in Figure 2 represents a randomized experiment with a compound treatment  $R$  and versions of treatment  $A$  (refer to the Appendix for a formal definition of  $A$  and of graphs). The graph does not include any common causes of the treatment  $R$  and the outcome  $Y$  because  $R$  is randomized. However, the graph includes common causes  $U$  of the versions  $A$  and the outcome  $Y$  because the versions  $A$  are not randomly assigned. For example, the actual duration of exercise  $A(1)$  chosen by those assigned to  $R = 1$  may depend on whether they have a history of heart disease  $L$ , which is the result of severe atherosclerosis  $U$ . Yet, if we let  $Y^r$  denote the outcome that would have resulted for an individual who was assigned to  $R=r$



FIGURE 1. Randomized experiment with a simple treatment  $A$ .

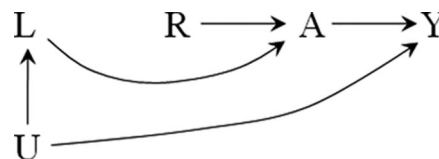


FIGURE 2. Randomized experiment with a compound treatment  $R$ .

and chose their version of treatment  $R=r$ , then the average causal effect  $\beta_R = E[Y^r] - E[Y^{r'}]$  is consistently estimated without any adjustment for covariates  $L$  because  $E[Y^r] = E[Y|R = r]$ .

The classification of treatments as either simple or compound is not straightforward. For example, we characterized 150 mg of aspirin as a simple treatment  $A$ , but suppose others argue that 150 mg of aspirin should be considered a compound treatment  $R$  because there are multiple versions of the 150-mg aspirin treatment. Two of these versions are (i) taking 150 mg of aspirin while holding the tablet with your left hand and (ii) taking 150 mg of aspirin while holding the tablet with your right hand. Arguably those 2 versions of treatment are irrelevant for the outcome because the hand with which the aspirin tablet is held will not affect an individual's survival. In this setting, we would make the assumption of treatment-variation irrelevance<sup>9</sup> for outcome  $Y$  (formally described below).

Under the assumption of treatment-variation irrelevance, compound treatments  $R$  can be safely viewed as simple treatments  $A$ . The assumption of treatment-variation irrelevance is often implicit in the interpretation of real data analyses, as all treatments are compound if one considers versions of treatment that are not relevant for the outcome of interest.

## COMPOUND TREATMENTS IN OBSERVATIONAL STUDIES

Consider an observational study to estimate the causal effect of a simple dichotomous non-time-varying treatment  $A$  on the 5-year risk of death  $Y$  (1: dead, 0: alive). We use the terms “treatment” and “exposure” interchangeably. As for the randomized experiment in the previous section, the treatment  $A$  may be taking a tablet of 150 mg of aspirin plus some particular combination of inactive ingredients ( $A = 1$ ) versus taking a tablet that is identical except it contains no aspirin ( $A = 0$ ), or it may be daily duration of exercise in a study population in which individuals exercise either 60 ( $A = 1$ ) or 10 ( $A = 0$ ) minutes per day. The average causal effect of treatment  $A$  on outcome  $Y$  is again defined as  $\beta_A = E[Y^a] - E[Y^{a'}]$ .

The causal graph in Figure 3 represents an observational study with a treatment  $A$ . The graph includes possibly unmeasured common causes  $U$  of the treatment  $A$  and the outcome  $Y$  because exchangeability is not guaranteed in observational studies. The graph also includes measured

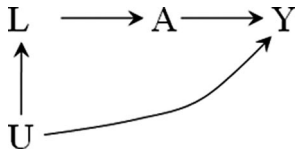


FIGURE 3. Observational study with a simple treatment  $A$ .

confounders  $L$  that are sufficient to block all backdoor paths between treatment  $A$  and outcome  $Y$ . That is, the graph represents a setting with exchangeability, or no unmeasured confounding, conditional on  $L$ . Under the conditions of exchangeability, positivity, and consistency (formally described in Appendix B), it is well known that  $\beta_A$  is consistently estimated via the standardized risk  $E[Y^a] = \sum_l E[Y|A = a, L = l] \Pr(L = l)$ .

Our latter example concerning exercise is, of course, artificial because one would not expect duration of exercise to be a dichotomous variable in a realistic study population. Rather, duration of daily exercise will take a range of values from 0 to, say, 180 minutes. Now suppose the investigators observe the daily minutes of exercise for each individual but decide to define a categorical variable  $R$  with 2 levels (1: “at least 30 minutes,” 0: “no more than 30 minutes”) for the analysis. Individuals who exercise longer than 30 minutes are classified as  $R = 1$ , and the others as  $R = 0$ . As in the randomized experiment described above, there are multiple versions of treatment  $A(r)$  within both categories  $R = 1$  and  $R = 0$ . Thus, when observational data are analyzed by defining categorical variables  $R$  for a continuous treatment, the variable  $R$  becomes a compound treatment. The decision to categorize the treatment variable is often made when the data are collected (eg, by asking participants to classify themselves in one of several categories rather than asking them their actual duration of exercise). The average causal effect of  $R$  on outcome  $Y$  is again defined as  $\beta_R = E[Y^r] - E[Y^{r'}]$ .

The graph in Figure 4 represents an observational study with a compound treatment  $R$ . In this graph, we assume that the actual duration of exercise  $A$  is the result of many nested decisions (to exercise more than 0 minutes, more than 5 minutes, more than 10 minutes ...), one of which (to exercise more than 30 minutes) is represented as the variable  $R$  in the

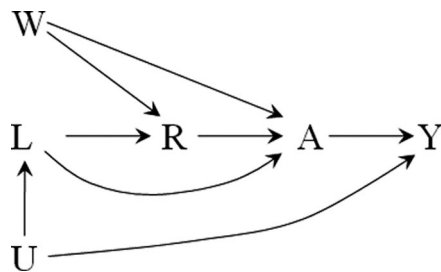


FIGURE 4. Observational study with a compound treatment  $R$  (treatment precedes versions).

graph. The graph includes an arrow from the confounders  $L$  to the decision  $R$ , and an additional set of confounders  $W$  to represent factors that affect both the decision  $R$  and the actual duration of exercise. Because the variables  $L$  and  $W$  are sufficient to block all backdoor paths between treatment  $R$  and outcome  $Y$  (including the path  $R \leftarrow W \rightarrow A \rightarrow Y$ ), the graph represents a setting with exchangeability, or no unmeasured confounding, conditional on  $\{L, W\}$ . Under exchangeability, positivity, and consistency (Appendix B),  $\beta_R$  is consistently estimated via the standardized risk

$$E[Y^r] = \sum_{l,w} E[Y|R = r, L = l, W = w] \Pr(L = l, W = w).$$

In observational studies, one can never guarantee exchangeability given  $L$  for simple treatments  $A$ , or  $\{L, W\}$  for compound treatments  $R$ . As a result, causal inference from observational data is controversial.

The notation  $E[Y^r]$  hides the fact that the magnitude of  $\beta_R$  depends on the versions of treatment  $A(r)$ . In the next paragraph, we rewrite  $\beta_R$  in a mathematically equivalent way that makes this dependence transparent. To do so, we need to expand our notation. First, let  $A_i^r(r)$  denote the counterfactual version of treatment  $R = r$  that subject  $i$  would receive if he received treatment level  $r$ . For example,  $A_i^1(1) = 43$  indicates that individual  $i$  would exercise 43 minutes if he exercised at least 30 minutes. Second, let  $Y_i^{r, A_i^r(r)}$  denote individual  $i$ 's counterfactual outcome if he received treatment value  $r$  by version  $A_i^r(r)$ . For example, the counterfactual outcome  $Y_i^{r=1, A_i^1(1)=37} = 1$  indicates that individual  $i$  would die if he exercised at least 30 minutes by exercising 37 minutes. Finally, let  $Y_i^{r, A_i^r(r)}$  denote individual  $i$ 's counterfactual outcome if he received compound treatment  $r$  under the version of treatment  $r$  that he would receive if he received treatment value  $r$ . For example,  $Y_i^{r=1, A_i^1(1)} = 0$  indicates that individual  $i$  would not die if he exercised at least 30 minutes by exercising for as long as he would if he exercised at least 30 minutes.

With this expanded notation, the average causal effect of the compound treatment  $R$  on the outcome  $Y$  can be redefined as  $\beta_R = E[Y^{r=1, A_i^1(1)}] - E[Y^{r=0, A_i^0(0)}]$ . This expanded notation clearly shows that the causal effect of the compound treatment  $R$  depends on the particular versions of treatment present in the population. It also helps clarify the consistency assumption for compound treatments  $R$ , as described in the next section.

### THE CONSISTENCY CONDITION FOR COMPOUND TREATMENTS

For a simple treatment  $A$ , the consistency condition is stated as  $Y_i = Y_i^a$  when  $A_i = a$  for all  $a$  and  $i$ . That is, consistency simply means that the death status for every treated individual in the study equals his death status if he had received treatment, and the death status for every untreated

individual in the study equals his death status if he had remained untreated. This statement seems obviously true.

For a compound treatment  $R$ , the consistency assumption for  $R$  effectively reduces to the consistency assumption for a simple treatment when the assumption of treatment-variation irrelevance<sup>9</sup> for outcome  $Y$  holds, ie, when

$$Y_i^{r,a(r)} = Y_i^{r,a'(r)} \text{ for all } r \text{ and } a(r), a'(r) \in \mathcal{A}(r).$$

Let us now consider the case in which the version of treatment is actually relevant to the outcome under consideration, that is, the condition of treatment-variation irrelevance does not hold. Then we may still articulate a consistency assumption as follows<sup>9</sup>: For individuals with  $R_i = r$ , we let  $A_i(r)$  denote the version of treatment  $R_i = r$ , actually received by individual  $i$ ; for individuals with  $R_i \neq r$ , we define  $A_i(r) = 0$  so that  $A_i(r) \in \{0\} \cup \mathcal{A}(r)$ . The consistency assumption would then require for all  $i$ ,

$$Y_i = Y_i^{r,a(r)} \text{ when } R_i = r \text{ and } A_i(r) = a(r).$$

That is, the death status for every individual in the study who received a particular version of treatment  $R = r$  (eg, 37 minutes of daily exercise) equals his death status if he had received that particular version of treatment. This statement is true by definition of version of treatment if we, in fact, define the counterfactual  $Y_i^{r,a(r)}$  for individual  $i$  with  $R_i = r$  and  $A_i(r) = a(r)$  as individual  $i$ 's outcome that he actually had under actual treatment  $r$  and actual version  $a(r)$ .

Suppose, for example, that the causal effect of exercising 37 minutes per day varies depending on whether an individual makes a spontaneous decision to exercise, or whether he is coerced to exercise. Then the definition of version of treatment  $A(r)$  could be expanded to include both the duration of exercise and the reason to exercise (spontaneous decision or coercion). Under this definition of versions of treatment, we can define counterfactuals so that consistency holds irrespective of the reason why individuals in our study exercised. Note that often observational studies collect data on treatments that have been spontaneously chosen by individuals, whereas counterfactuals are often defined in terms of interventions that may involve some degree of coercion or another form of influence. In such settings with counterfactuals defined by specific interventions, the consistency statement then essentially assumes treatment-variation irrelevance.<sup>9</sup> We return to this point in the discussion.

The consistency condition links the observed data to the counterfactual outcomes. In the absence of consistency, one would not know which counterfactual contrast is being estimated by the data. That is, one would not know which causal effect  $\beta_R$ , if any, is being estimated, and thus it would be difficult to justify the use of the effect estimates for any sort of decision making. However, by definition, one can assume that some form of the consistency condition always holds

regardless of whether the treatment is simple or compound and regardless of whether the version of treatment is relevant to the outcome of interest.

Therefore, if exchangeability and positivity also hold, we will be able to validly estimate the average causal effect  $\beta_R$  from observational data collected in a particular population. A question that then arises is whether  $\beta_R$  will vary across populations.

## TRANSPORTABILITY OF THE CAUSAL EFFECT OF COMPOUND TREATMENTS

Causal effects estimated in one population (the study population) are often intended for use in making decisions in another population (the target population). Suppose we have correctly estimated the average causal effect  $\beta_R$  of compound treatment  $R$  in our study population, but we want to know the average causal effect of compound treatment  $R$  in a different target population. Can we say that the effect in the target population is the same as in the study population? That is, can we “transport”—or extrapolate or generalize—the effect from one population to the other? This is a question of external validity. The answer to this question depends on the characteristics of both populations. Specifically, transportability of effects from one population to another may be justified if the following characteristics are similar between the 2 populations:

1. Effect modification: The effect of treatment may differ across individuals with different susceptibility to the outcome. For example, if women are more susceptible to the effects of exercise than men, we say that sex is an effect modifier for exercise. The distribution of effect modifiers in a population will generally affect the magnitude of the causal effect of treatment in that population. Discussions about transportability of causal effects have often been restricted to effect modification.
2. Interference: In many settings, treating one individual may indirectly affect the treatment level of other individuals in the population. For example, a socially and physically active individual may convince his friends to exercise with him, and thus an intervention on that individual may be more effective than an intervention on a socially isolated individual. The distribution of contact patterns among individuals may affect the magnitude of the causal effect in a population. The relevance of this interference for transportability of effects is increasingly being recognized.<sup>13–16</sup>
3. Versions of the compound treatment: To see why the versions of treatment may affect transportability, suppose that all individuals who exercise at least 30 minutes/day do so for 60 minutes/day in the first population and for 31 minutes/day in the second one. Then, if the duration of exercise is relevant to the outcome of interest, the effect of exercising at least 30 minutes/day versus less than 30 minutes/day will generally differ between the 2 popula-



tions, even if their distributions of effect modifiers and interference patterns are identical.

The causal effect of a compound treatment  $R$  depends on the distribution of versions  $A(r)$  of treatment  $R$  in the study population. The mean counterfactual outcome under treatment level  $R = r$ ,  $E[Y^{r, A(r)}]$ , in a particular population can be interpreted as the mean counterfactual outcome under a random treatment regimen in which subjects with covariate values  $L = l$ ,  $W = w$  are assigned to version  $A(r) = a(r)$  with probability  $\Pr[A(r) = a(r)|r, l, w]$ . Thus, the average causal effect of the compound treatment (or regimen)  $R$  will differ between 2 populations with a different set of probabilities  $\Pr[A(r) = a(r)|r, l, w]$ . That is, the effect of exercising at least 30 minutes/day in the study population may be inappropriate to make decisions about public health policy in the target population, even if the distribution of effect modifiers and interference patterns are identical between the 2 populations.

### WHEN THE VERSIONS OF THE COMPOUND TREATMENT ARE KNOWN

There is one potential solution to the transportability problem. One could collect data on the version of treatment  $A(r)$  received by each individual in the study population, and then choose a set of probabilities  $\Pr^*[A(r) = a(r)|R = r, L = l, W = w]$  to estimate the average causal effect of compound treatment  $R$  under the relevant distribution of versions of treatment. Specifically, the counterfactual mean of the outcome under the regimen  $R$  in the study population is<sup>12</sup>

$$\begin{aligned} E[Y^r] &= E[Y^{r, A(r)}] \\ &= \sum_{l, w, a(r)} E[Y | A(r) = a(r), L = l, W = w] \\ &\quad \Pr[A(r) = a(r)|r, l, w] \Pr(l, w). \end{aligned}$$

In the above equation, one could replace the probabilities  $\Pr[A(r) = a(r)|r, l, w]$  from the study population by the set of probabilities  $\Pr^*[A(r) = a(r)|r, l, w]$  from the target population to estimate the effect of the compound treatment  $R$  in the target population, as long as all the versions of treatment present in the target population are also present in the study population, that is,  $\Pr[A(r) = a(r)|r, l, w] > 0$  with probability 1 for all  $A(r)$  in the target population. Alternatively, one can choose the probabilities  $\Pr^*[A(r) = a(r)|r, l, w]$  to represent those under an intervention of interest. For example, if a public health program to encourage at least 30 minutes of daily exercise is hypothesized to result in 60% of the population exercising 30 minutes, 20% exercising 40 minutes, etc, these probabilities can be used in the above expression to estimate the effect of the program. Policy makers could estimate the effect under different hypothesized distributions of duration of exercise as a sort of sensitivity analysis. As discussed in Appendix A, the same relation between versions of treatment in an observational setting and as implemented in a specific policy holds even if it is thought

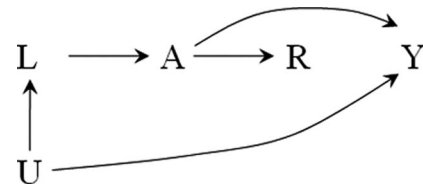


FIGURE 5. Observational study with a compound treatment  $R$  (versions precede treatment).

that the variable  $R$  is just an investigator-created deterministic function of duration of exercise  $A$  and thus  $A$  should precede  $R$  as shown in Figure 5.

Unfortunately, reconciling different distributions of version of treatments across populations is not always straightforward. Besides the possibility that some versions of treatment may not exist in both populations, there are 2 main obstacles to implement a strategy based on replacing the set of probabilities  $\Pr[A(r) = a(r)|r, l, w]$  with  $\Pr^*[A(r) = a(r)|r, l, w]$ .

First, there may be too many versions  $A(r)$  of compound treatment  $R$ . Take the treatment “exercise at least 30 min/d.” We have so far made the oversimplifying assumption that all relevant versions of this treatment can be put in one-to-one correspondence with an indexing discrete variable, that is, the duration of daily exercise in minutes. However, it is likely that the type (eg, swimming, biking, running) and intensity (mild, moderate, strenuous) of exercise are also relevant to the outcome. The number of versions of treatment increases substantially when one considers all possible combinations of duration, type, and intensity of exercise. Even more versions of treatment can be defined if one considers weekly patterns, equipment, time of day, degree of coercion, etc. One obvious problem is the difficulty of enumerating all versions of treatment, which would be necessary to adopt the above strategy to transport effect estimates. Unless the assumption of treatment-variation irrelevance holds for many of these versions (eg, same effect under different degrees of coercion, time of day, etc), the transportability of causal inferences will be impractical.

Second, the versions of treatment may be unmeasured or unknown. The next section discusses the consequences of this problem for the transportability and identification of causal effects. It could be argued that this problem might not occur for the versions of the treatment “exercise at least 30 minutes daily” because, as many as they may be, the versions could all be enumerated by experts in the field. Therefore, in the next section, we switch to another example of a common treatment (or exposure) in epidemiology: change in body weight.

### WHEN THE VERSIONS OF THE COMPOUND TREATMENT ARE UNKNOWN

A debate exists on how many deaths would be prevented every year in the United States if overweight (defined as a body mass index [BMI] between 25 and 30) and obesity

(defined as BMI >30) were eliminated. That is, the debate revolves around the effect  $\beta_R$  of the compound treatment  $R$ , ie, “BMI >25” (1: yes, 0: no). Epidemiologic studies try to estimate  $\beta_R$ , and thus the number of excess deaths attributable to overweight and obesity, by comparing the mortality risk in people at different levels of BMI. Suppose a study conducted in a random sample of the US population estimated that 100,000 deaths/year are attributable to  $R = 1$ ; that is, 100,000 deaths would be avoided (or, rather, postponed) under an intervention that eliminates overweight and obesity. The question is: what intervention is that?

Many versions of  $R = 1$  (ie, BMI >25) are present in the US population, including any combination of low physical activity, high caloric intake, no cigarette smoking, low basal metabolic rate, certain genetic factors and gastrointestinal bacteria, lack of bariatric surgery, and many others, including those yet unknown. Some of the versions whose existence is known, such as genetic factors, are poorly understood. Thus, the estimate of 100,000 excess deaths due to  $R = 1$  estimates the effect of a treatment regimen  $R$  that assigns each individual to one of the multiple—and possibly unknown—versions  $A(r)$  of treatment  $R$  in the US population, with a probability that may depend on the individual’s characteristics  $L$  and  $W$ . Since the versions of treatment are partly unknown, we can neither enumerate them nor describe their distribution. It follows that we do not know which causal effect  $\beta_R$  is being estimated and thus the estimate of excess deaths is not necessarily meaningful for public policy.

Specifically, suppose a policy-maker wishes to know the expected effect of an obesity prevention program if it were nationally implemented in the United States. The prevention program under consideration would eliminate low physical activity and high caloric intake, but would not affect any other factors such as an individual’s basal metabolic rate (which is tightly controlled by genetic and other factors). Can the effect estimate from our study population be directly translated into the effect of the proposed prevention program? Not in general. Though the distribution of effect modifiers and the interference patterns in our study population are expected to equal those in the US population (because our study population is a representative sample of the US population), our effect estimate corresponds to a treatment regimen that is different from the treatment regimen considered under the obesity prevention program. For example, consider  $a(r = 0) =$  “very high basal metabolic rate,” a version of the treatment  $r = 0$  or “BMI  $\leq 25$ .” For this version  $\Pr[A(0) = a(0)|R = 0, L, w] > 0$  in the regimen operating in our study population, but  $\Pr^*[A(0) = a(0)|R = 0, L, w] = 0$  in the regimen considered under the obesity prevention program. Thus the actual number of deaths avoided by the intervention could be, say, 30,000 if the exercise/diet regimen is less effective than some regimes operating in the population; alternatively it could be 300,000 if more effective. The use of

the estimate from observational data to predict the magnitude of the effect of the prevention program would be justified only under the assumption of treatment-variation irrelevance for all versions of treatment.

Another problem arises when the versions  $A(r)$  of treatment  $R$  are unknown. The identification of the average causal effect of compound treatment  $R$  requires exchangeability and positivity, in addition to consistency. Exchangeability holds if the covariates contain a sufficient set of variables to control for confounding. The choice of the covariates needs to be guided by subject matter knowledge.<sup>17</sup> However, if the versions of treatment for compound treatment  $R$  are not explicitly defined, the identification and measurement of the relevant covariates becomes a daunting task. For example, suppose genetic polymorphism  $b$  is one of the versions of treatment  $R$ , and genetic polymorphism  $c$  is a cause of  $R$  that shares a common cause with  $b$ . How can an investigator know that it is necessary to adjust for  $c$ , not knowing that  $b$  is one of the versions of treatment  $R$ ? By not providing a full characterization of the versions of treatment, one is not only estimating a vague causal effect but also reducing the likelihood of estimating a vague causal effect without bias. Moreover, suppose that investigators happened to measure and adjust for  $c$  (even though they did not know that  $b$  was a version of treatment), and that  $c$  is in strong linkage disequilibrium with  $b$ . Then it is possible that the positivity condition will not hold because the probability of having  $c$  is 0 for those lacking  $b$ .

Thus, unknown versions of treatment imply the impossibility of evaluating the exchangeability and positivity conditions.<sup>7</sup> Though collecting information on the versions of treatment is not necessary to estimate the effect of the compound treatment  $R$  in the study population (see previous section), this information is necessary to transport the effect to other populations or to evaluate the effect of well defined intervention programs. We cannot use our expert knowledge to identify and measure confounders for versions of treatment if we do not know what the versions of treatment are.

## DISCUSSION

This article continues a recent thread of papers, initiated by Robins and Greenland,<sup>3</sup> on the implications of ill-defined counterfactuals for causal inference. Hernán<sup>5</sup> related the issue to the existence of multiple versions of treatment, and expressed concern about overreliance on sophisticated statistical methods for causal inference in the absence of well-defined causal questions. Later Hernán and Taubman<sup>7</sup> extended the discussion on multiple versions of treatment and explicitly linked ill-defined causal questions to departures from the consistency condition. Using the epidemiologic exposure “obesity” as an example, they described the causal question that epidemiologists implicitly ask when estimating the causal effect of obesity, and concluded that such a

question is not guaranteed to be relevant for public health decision-making. They argued further that asking such causal question makes it difficult to appropriately measure and adjust for all confounders. The link between consistency and multiple versions of treatment was further explored by Cole and Frangakis,<sup>8</sup> and by VanderWeele,<sup>9</sup> who provided a formal definition of consistency in the presence of multiple versions of treatment. Joffe et al<sup>18</sup> considered the estimation of effects of a compound treatment under the assumption of treatment-variation irrelevance.

All the above authors seem to agree that caution is needed when multiple (relevant) versions  $A(r)$  of treatment  $R$  exist because, in this setting, the counterfactual outcomes  $Y^r = Y^{r, A^r(r)}$  are vaguely defined. This logically entails the concern that consistency cannot be taken for granted in observational studies. Other authors, however, do not share this view.

van der Laan et al<sup>6</sup> acknowledge the existence of multiple versions of treatment and the impossibility of enumerating them, but conclude that “formally, the existence of the counterfactuals is a nonissue” and that the parameter  $\beta_R$  is “a very interesting one [...] without concern as to well-defined outcomes for the given levels of treatment.”

Pearl<sup>10</sup> argues that the consistency condition is a mathematical theorem that always holds true, regardless of the definition of treatment, as long as the causal model is correct. Specifically, his causal model is a nonparametric structural equation model represented by a causal directed acyclic graph. To Pearl, consistency is not a condition that might always hold true by definition of counterfactual outcomes and versions of treatments (as in our exposition above), but rather the logical consequence of using that particular causal model. He also emphasizes that the discussion on whether the consistency condition should be referred to as a “theorem” or an “assumption” depends on the causal framework; it is a theorem under the causal directed acyclic graph framework but generally an assumption under the potential outcomes framework. He also argues that the distinction goes beyond semantics, because “theorem conveys to practitioners the comfortable presence of a solid science behind their practice and the assurance that this science can be relied upon for guidance despite its dealing with ideal mathematical objects.” Pearl goes further in declaring that the consistency condition holds true for any variables in the causal model, including any covariates  $L$  and  $W$ .

Thus, a debate has ensued as to whether the consistency condition should be considered as (i) an assumption that one needs to evaluate; (ii) an axiom, something to be taken for granted, that requires no further evaluation; or (iii) a theorem that follows directly from the causal model. We claim that all 3 positions are valid, albeit from different standpoints.

The identification of the average causal effect  $\beta_R$  of compound treatment  $R$  requires the compound consistency con-

dition  $Y_i = Y_i^{r, a(r)}$  and  $A_i(r) = A_i^r(r)$ , when  $R_i = r$  and  $A_i(r) = a(r)$  for all individuals  $i$ . Interestingly, when the version  $A_i(r)$  of treatment  $R = r$  that individual  $i$  receives is unknown, the consistency condition cannot be articulated for individual  $i$ . When faced with this logical problem, at least 3 responses are possible corresponding to how one interprets the consistency assumption.

First, one could argue that the consistency condition can be well characterized only under the assumptions that all version of treatments  $A(r)$  are known. That is, consistency is a substantive assumption that needs to be evaluated.

Second, one could argue that knowing the particular versions  $A(r)$  is irrelevant: the consistency condition will hold for individual  $i$  by definition, whether we do or do not know his version of treatment, for some unspecified version of treatment  $A_i(r)$ . Thus, consistency is an axiom related to the definition of counterfactual outcomes. This response is problematic because if one renounces the task of characterizing the versions of treatment, one is automatically waiving the right to characterize the causal effect  $\beta_R$  that is being estimated. If one does not know the distribution of versions of treatments that corresponds to the causal effect that has been estimated, then one does not know whether a well-defined intervention program will have an effect that bears much resemblance to the estimate that had been obtained (even if the exchangeability, consistency, and positivity assumptions hold). Under this second perspective, the consistency condition may never be strictly violated in observational studies but, in the presence of unknown multiple versions of treatment, the consistency condition is used to estimate the effect  $\beta_R$  of a treatment regimen  $R$  that (i) is population-specific and undefined and (ii) may not be possible to implement in practice.

Third, under a nonparametric structural equation model, consistency is logically inferred from the properties of the model. Thus, consistency is a theorem, though one for which its application requires that the nonparametric structural equation model represented on a causal diagram be correct. This third response is closely related to the first response: the specification of a causal diagram representing a nonparametric structural equation model that adequately captures the different versions of treatment requires an adequate characterization of the versions of treatment. Consistency is guaranteed from the causal diagram if the causal diagram is a correct model, but unless the causal diagram incorporates the various versions of treatment, the causal diagram may not be correct.

Although all 3 perspectives are to some extent valid, we believe that investigators will be best served by the first perspective, which is the most relevant for policy- and decision-making. However, irrespective of how one interprets the consistency condition, if inadequate attention is given to



evaluating the condition or to explicitly modeling versions of treatment on a causal diagram, one runs the risk of making causal inferences that cannot appropriately inform decision making. Actionable effect estimates require knowledge of versions of treatment.

The problems discussed in this paper apply to many common epidemiologic exposures that can be conceptualized as having multiple relevant versions. Some examples, in addition to physical activity and body weight, are nutrients, biomarkers, and socioeconomic status. Again, the simple-compound dichotomy is artificial: few, if any, epidemiologic exposures are truly simple treatments with no relevant versions. However, the vagueness in the definition of causal effects is more tolerable for some treatments (eg, 150 mg/day of aspirin) than for others (eg, BMI >25).

For simplicity, throughout this paper we have ignored that most treatments are time-varying. Causal inference for time-varying treatments requires the consideration of additional issues such as the timing of the intervention and the fact that the version of treatment at time  $t$  may be a strong confounder for the compound treatment at subsequent times. In fact, under the ordering of variables represented in Figure 4,  $A_i(r)$  measured at time  $t$  can be viewed as part of  $L_{t+1}$  for all practical purposes.

Several authors have discussed time-varying compound treatment or regimes. Taubman et al<sup>19</sup> considered threshold and representative treatment regimes that differ only with respect to the probabilities  $\Pr[A_i(r) = a(r) | R = r, L_t = 1]$ . Cain et al<sup>20</sup> considered dynamic treatment regimes that differ only with respect to the probability of treatment initiation  $\Pr[A_i(r) = a(r) | R = r, L_t = 1]$  during a fixed time period after certain threshold is reached. Robins et al<sup>21</sup> and Hernán et al<sup>22</sup> provided formulations of versions of treatment that include the patterns of observation and measurement. VanderWeele and Vansteelandt<sup>23</sup> discuss the consistency assumption in the context of direct and indirect effects with treatment variables at different times in which the second treatment variable is taken as the mediator.

In summary, investigators conducting observational studies are typically, and rightly, concerned about confounding, measurement error, and selection bias. But the emphasis on these potential biases often obscures a central issue in causal inference: the specification of the causal question. In the absence of a well-defined question, one does not know which causal effect, if any, is being estimated, and thus it becomes harder to justify the use of the resulting effect estimate for decision-making.

## APPENDIX A: DIRECTED ACYCLIC GRAPHS WITH COMPOUND TREATMENTS

Recall  $A_i(r)$  denotes the version of treatment  $R = r$  that individual  $i$  received and is in some set  $\mathcal{A}(r)$ ;  $A_i^*(r)$  denotes the counterfactual version of treatment  $R = r$  that subject  $i$

would receive if he received treatment level  $r$ ; also  $Y_i^{r,a(r)}$  denotes individual  $i$ 's counterfactual outcome if he received treatment value  $r$  by version  $a(r)$ , so that  $Y_i^{r,A_i^*(r)}$  denotes individual  $i$ 's counterfactual outcome if he received treatment value  $r$  under the version of treatment  $r$  that he would receive if he received treatment value  $r$ . This notation presupposes that, for a subject with  $R_i = r$ , the counterfactuals  $A_i^*(r^*)$ , with  $r^* \neq r$ , are well defined, and that for a subject with  $R_i = r$  and  $A_i(r) = a(r) \in \mathcal{A}(r)$ , (i) the counterfactual outcomes  $Y_i^{r,a'(r)}$  with  $a'(r) \neq a(r)$ , and (ii) the counterfactual outcomes  $Y_i^{r^*,a(r^*)}$  with  $r^* \neq r$ ,  $a(r^*) \in \mathcal{A}(r^*)$  are well defined.

In Figures 2 and 4, we define  $A_i = (I[A_i(r) = a(r)] : r \in R, a(r) \in \mathcal{A}(r))$  as a  $\sum_r n(r)$ -dimensional vector of indicators (ie, dichotomous variables)  $I[A_i(r) = a(r)]$  for each version  $a(r)$  of each level  $r$  of compound treatment  $R$ . For individual  $i$ , all the entries in vector  $A_i$  are 0, except the entry corresponding to the version  $A_i(r) = a(r)$  that individual  $i$  actually received. This definition of  $A$  implies that the causal effect of  $R$  on  $Y$  is entirely mediated through the actual versions of treatment  $A$ , ie, that there is no direct arrow from the compound treatment  $R$  to the outcome  $Y$ .

The following example helps describe the difference between the definition of  $A$  as a simple treatment in Figures 1 and 3, and as a vector of versions of treatment in Figures 2 and 4. Consider again the randomized trial in which patients are assigned to either  $R = 1$  "exercise at least 30 minutes daily" or  $R = 0$  "exercise less than 30 minutes daily." Because the assignment cannot be blinded, it is possible that many participants assigned to  $R = 1$  feel less guilty about eating large quantities of food and therefore increase their caloric intake beyond their increased caloric expenditure. If  $A$  stands for the duration of daily exercise, we would say that  $R$  has a direct effect on  $Y$  that is not mediated by  $A$ , and would add a direct arrow from  $R$  to  $Y$  to the causal graph. However, if as in Figures 2 and 4,  $A$  stands for the versions of treatment  $R$ , one of those versions would include "increased caloric intake because of being assigned to  $R = 1$ ," and thus, we add no direct arrow from  $R$  to  $Y$ .

Figure 4 does not always represent the appropriate temporal relation between the compound treatment  $R$  and the versions of treatment  $A$  in observational studies. For example, the value of BMI may be viewed as the result of the action of different factors (eg, genes, diet, exercise, intestinal flora, etc) that define the versions of treatment, which would imply that the versions  $A$  should precede  $R$  in the causal graph. Figure 5 depicts this setting in which the direction of the arrow from  $R$  to  $A$  is reversed and, because  $R$  is fully determined by  $A$ , there are no direct arrows from any nodes other than  $A$  into  $R$ . That is,  $R$  and  $A$  do not share any common causes  $W$ .

The choice between Figure 4 and Figure 5 is not always obvious in observational studies. For example, we decided to



use Figure 4 to depict the dichotomous treatment  $R$ , with  $R = 1$  being “exercise at least 30 minutes,” and  $R = 0$  being “no more than 30 minutes.” Our rationale for this choice was that the actual duration of exercise  $A$  can be thought of as the result of many nested decisions (to exercise more than 0 minutes, more than 5 minutes, more than 10 minutes, etc), one of which (to exercise more than 30 minutes) is represented as the variable  $R$  in the graph. Under this conceptualization,  $R$  precedes  $A$  as shown in Figure 4. However, one could also argue that the variable  $R$  is just an investigator-created deterministic function of duration of exercise  $A$ ; and thus,  $A$  should precede  $R$  as shown in Figure 5.

It can, however, be shown<sup>12</sup> that under Figure 5, the same comments hold regarding the relation between what is estimated in an observational study and the effect one would expect from a particular intervention. Specifically, the effect estimate in an observational study comparing  $R = 1$  and  $R = 0$ , adjusted for  $L$ , in Figure 5, can be interpreted as a comparison in a randomized trial in which, within strata of covariates  $L = l$ , individuals in one arm are randomly assigned a “version of treatment”  $A$  from the observed distribution of  $A$  in the population among those with  $R = 1$  and  $L = l$ , and individuals in the other arm are randomly assigned a “version of treatment”  $A$  from the observed distribution of  $A$  in the population among those with  $R = 0$  and  $L = l$ .

Nevertheless, several points concerning interpretation merit attention with respect to Figure 5. First, no adjustment for  $W$  is necessary because no common causes  $W$  exist in Figure 5. Second, the average causal effect of  $R$  itself on outcome  $Y$ ,  $\beta_R = E[Y^r] - E[Y^{r'}]$ , is by definition null, because there is no causal pathway from  $R$  to  $Y$  in Figure 5. The lack of direct effect of  $R$  on  $Y$  (ie, no direct arrow from  $R$  to  $Y$ ) is common to Figures 4, 5; but in Figure 4, a non-null average causal effect of  $R$  will exist if the versions  $A$  have an effect on the outcome, ie, through the pathway  $R \rightarrow A \rightarrow Y$ . In contrast, in Figure 5,  $R$  has a null average causal effect even if the versions  $A$  have an effect on the outcome, ie, the pathway  $R \leftarrow A \rightarrow Y$  does not represent an effect of  $R$  mediated through  $A$ . As a consequence, the estimate  $\hat{\beta}_R$  under Figure 5 is not really quantifying the average causal effect of  $R$  (which is by definition null) but rather the association between  $R$  and  $Y$  that is mediated through their common cause  $A$ . As noted in the previous paragraph, this effect estimate  $\hat{\beta}_R$  in Figure 5 can be interpreted as an estimate of what would have been observed in the randomized trial described above. Third, in this setting,  $R$  acts as a mismeasured form of the truly causal variable  $A$ , as discussed by Hernán and Cole.<sup>24</sup> Thus Figure 5 is an even stronger reminder that one needs to characterize the versions of treatment before giving a causal interpretation to associations estimated from observational data.

## APPENDIX B: IDENTIFIABILITY ASSUMPTIONS

It is well known that, for a simple treatment  $A$ , the average causal effect  $\beta_A$  is consistently estimated via the standardized risk  $E[Y^a] = \sum_l E[Y|A = a, L = l] \Pr(L = l)$  when the following identifiability conditions hold for all  $a$ ,

$$Y^a \perp\!\!\!\perp A|L \text{ (exchangeability)}$$

$$\Pr[A = a|L] > 0 \text{ if } f_L(l) > 0 \text{ (positivity)}$$

$$Y_i = Y_i^a \text{ when } A_i = a \text{ for all individuals } i, \text{ (consistency)}$$

where  $L$  is a vector of (discrete) covariates. If  $L$  contains continuous covariates, then  $E[Y^a] = \int_l E[Y|A = a, L = l] dF_L(l)$ .

For a compound treatment  $R$  with multiple (relevant) versions  $A(r)$ , VanderWeele and Hernán<sup>12</sup> show that the average causal effect  $\beta_R$  is consistently estimated via the standardized risk

$$\begin{aligned} E[Y^r] &= E[Y^{r, A(r)}] \\ &= \sum_{l,w} E[Y|R = r, L = l, W = w] \\ &\quad \Pr(L = l, W = w) \end{aligned}$$

when the following identifiability conditions hold for all  $r$  and  $a$ ,

$$Y^r \perp\!\!\!\perp R \{L, W\} \text{ (compound exchangeability)}$$

$$\Pr[R = r|L = l, W = w] > 0 \text{ if } f_{L,W}(l, w) > 0 \text{ (compound positivity)}$$

$$\text{For all individuals } i, Y_i = Y_i^{r, a(r)} \text{ and } A_i(r) = A_i^r(r)$$

$$\text{when } R_i = r \text{ and } A_i(r) = a(r) \text{ (compound consistency)}$$

where  $L$  and  $W$  are 2 vectors of measured covariates with  $W$  denoting variables that only affect treatment and version and not the outcome directly.

Note that only the consistency condition needs further elaboration for compound treatments compared with simple treatments. The exchangeability and positivity conditions for a compound treatment  $R$  are identical to the exchangeability and positivity conditions for a simple treatment  $A$ , respectively, with  $A$  replaced by  $R$ ,  $a$  by  $r$ , and  $L$  by  $(L, W)$ .

For simplicity, and without loss of generality, we restricted our discussion to discrete covariates. Also note that the above standardized risks could also be expressed as inverse probability weighted risks.

## ACKNOWLEDGMENTS

We thank Stephen Gilman, Sonia Hernández-Díaz, and Marshall Joffe for helpful comments.

## REFERENCES

- Lewis D. *Counterfactuals*. Oxford: Blackwell; 1973.
- Holland PW. Statistics and causal inference. *J Am Stat Assoc*. 1986;81:495–970.
- Robins JM, Greenland S. Comment on “Causal inference without counterfactuals.” *J Am Stat Assoc*. 2000;95:477–482.
- Greenland S. Causality theory for policy uses of epidemiologic measures. In: Murray CJ, Salomon JA, Mathers CD, et al, eds. *Summary Measures of Population Health*. Cambridge, MA: Harvard University Press/World Health Organization; 2002.
- Hernán MA. Invited commentary: hypothetical interventions to define causal effects: afterthought or prerequisite? *Am J Epidemiol*. 2005;162:618–620.
- van der Lan MJ, Haight TJ, Tager IB, van der Lan et al. respond to “Hypothetical interventions to define causal effects.” *Am J Epidemiol*. 2005;162:621–622.
- Hernán MA, Taubman SL. Does obesity shorten life? The importance of well-defined interventions to answer causal questions. *Int J Obes*. 2008;32(suppl 3):S8–S14.
- Cole SR, Frangakis CE. The consistency statement in causal inference: A definition or an assumption? *Epidemiology*. 2009;20:3–5.
- VanderWeele TJ. Concerning the consistency assumption in causal inference. *Epidemiology*. 2009;20:880–883.
- Pearl J. On the consistency rule in causal inference: an axiom, definition, assumption, or a theorem? *Epidemiology*. 2010;21:872–875.
- VanderWeele TJ, Hernán MA. Causal effects and natural laws: towards a conceptualization of causal counterfactuals for non-manipulable exposures with application to the effects of race and sex. In: C. Berzuini, P. Dawid, and L. Bernardinelli (Eds.). *Causality: Statistical Perspectives and Applications*. Wiley, 2011. In press.
- VanderWeele TJ, Hernán MA. Causal inference under multiple versions of treatment (February 2011). *COBRA Preprint Series*. Article 77. <http://biostats.bepress.com/cobra/ps/art77>.
- Halloran ME, Struchiner CJ. Causal inference in infectious diseases. *Epidemiology*. 1995;6:142–151.
- 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.
- Rosenbaum PR. Interference between units in randomized experiments. *J Am Stat Assoc*. 2007;102:191–200.
- Hudgens MG, Halloran ME. Towards causal inference with interference. *J Am Stat Assoc*. 2008;103:832–842.
- Hernán MA, Hernández-Díaz S, Werler MM, Mitchell AA. Causal knowledge as a prerequisite for confounding evaluation: an application to birth defects epidemiology. *Am J Epidemiol*. 2002;155:176–184.
- Joffe M, Wei Y, Brunelli S, Feldman H. Methodological issues in the study of the effects of hemoglobin variability. *UPenn Biostatistics Working Papers*. 2007: Paper 19. Available at :<http://biostats.bepress.com/upennbiostat/papers/art19>.
- Taubman SL, Robins JM, Mittleman MA, Hernán MA. Alternative approaches to estimating the effects of hypothetical interventions. In: *JSM Proceedings 2008, Health Policy Statistics Section*. Alexandria, VA: American Statistical Association.
- Cain LE, Robins JM, Lanoy E, Logan R, Costagliola D, Hernán MA. When to start treatment? A systematic approach to the comparison of dynamic regimes using observational data. *Int J Biostat*. 2010;6:Article 18. Available at: <http://www.bepress.com/ijb/vol6/iss2/18>.
- Robins JM, Orellana L, Rotnitzky A. Estimation and extrapolation of optimal treatment and testing strategies. *Stat Med*. 2008;27:4678–4721.
- Hernán MA, McAdams M, McGrath N, Lanoy E, Costagliola D. Observation plans in longitudinal studies with time-varying treatments. *Stat Methods Med Res*. 2009;18:27–52.
- VanderWeele TJ, Vansteelandt S. Conceptual issues concerning mediation, interventions and composition. *Stat Interface*. 2009;2:457–468.
- Hernán MA, Cole SR. Causal diagrams and measurement bias. *Am J Epidemiol*. 2009;170:959–962.