Estimation of Direct Causal Effects

Maya L. Petersen, Sandra E. Sinisi, and Mark J. van der Laan

Abstract: Many common problems in epidemiologic and clinical research involve estimating the effect of an exposure on an outcome while blocking the exposure's effect on an intermediate variable. Effects of this kind are termed direct effects. Estimation of direct effects is typically the goal of research aimed at understanding mechanistic pathways by which an exposure acts to cause or prevent disease, as well as in many other settings. Although multivariable regression is commonly used to estimate direct effects, this approach requires assumptions beyond those required for the estimation of total causal effects. In addition, when the exposure and intermediate variables interact to cause disease, multivariable regression estimates a particular type of direct effect—the effect of an exposure on an outcome when the intermediate is fixed at a specified level. Using the counterfactual framework, we distinguish this definition of a direct effect (controlled direct effect) from an alternative definition, in which the effect of the exposure on the intermediate is blocked, but the intermediate is otherwise allowed to vary as it would in the absence of exposure (natural direct effect). We illustrate the difference between controlled and natural direct effects using several examples. We present an estimation approach for natural direct effects that can be implemented using standard statistical software, and we review the assumptions underlying our approach (which are less restrictive than those proposed by previous authors).

(*Epidemiology* 2006;17: 276–284)

Many research questions in epidemiology are concerned with understanding the causal pathways by which an exposure or treatment affects an outcome. Research may be aimed at understanding the mechanisms by which a treatment or exposure affects disease. For example, in HIV-infected individuals, protease inhibitor-based antiretroviral therapy preserves CD4 T-cell counts. Are these beneficial effects due entirely to reductions in plasma HIV RNA level (viral load)? Alternatively, in settings where individuals and their physicians alter their treatment decisions as a result of exposure, research may be aimed at estimating the causal effects of an exposure if there were no effect of the exposure on the treatment decision. For example, air pollution can affect children's lung function. However, high pollutant levels may also cause children to increase their use of rescue medication. How would an increase in pollutant levels affect lung function if medication use did not increase?

These are 2 examples of a common causal structure underlying epidemiologic problems. This structure can be represented in the form of a directed acyclic graph (DAG) (Fig. 1).[1] In the applications considered in this article, the exposure of interest acts on the outcome through 2 pathways: one in which the exposure affects an intermediate variable which in turn affects the outcome, and one in which effects of the exposure are not through changes in the intermediate. In these examples, the goal is to estimate direct effects, ie, effects of the exposure on the outcome if the exposure's effect on the intermediate variable was blocked.

Epidemiologists and others have generally used standard analytic approaches such as multivariable regression to estimate direct effects. Such approaches can provide a reasonable test of the null hypothesis that no direct effect is present; however, the validity of this approach relies on several assumptions which, although raised previously, may not be widely appreciated.[2–6] In cases in which the necessary assumptions are met and exposure and intermediate interact to cause disease, multivariable regression can estimate the direct effect of an exposure at each controlled level of the intermediate variable (the controlled direct effect).[3] Alternatively, a direct effect can be defined as the effect of an exposure on an outcome, blocking only the effect of the exposure on the intermediate (natural direct effect).[2,3] Note that, in the former definition, all causal effects on the intermediate are blocked, whereas in the latter case, only the effect of the exposure on the intermediate is blocked (Figs. 2A and 2B, respectively). In this article, we discuss these 2 distinct types of direct effects—controlled and natural, which can be used to answer distinct research questions. Controlled direct effects can be estimated directly from multivariable regression. When exposure and intermediate interact, as many controlled direct effects exist as there are levels of the intermediate. In this setting, estimation of the natural direct effect, which provides a summary of the direct effect of exposure in the population, requires additional steps.

The article is structured as follows. First, we review the definition and interpretation of controlled and natural direct effects using examples and the counterfactual framework for causal inference. Then we present a simple method for the estimation of natural direct effects (a numeric illustration is

Submitted 17 November 2004; accepted 16 November 2005.
From the Division of Biostatistics, University of California, Berkeley, CA.
Maya Petersen is supported by a Predoctoral Fellowship from the Howard Hughes Medical Institute. Mark van der Laan is supported by NIMH grant R01 GM071397.
Correspondence: Maya L. Petersen, University of California at Berkeley, Division of Biostatistics, School of Public Health, Earl Warren Hall #7360, Berkeley, California 94720-7360. E-mail: mayaliv@ berkeley.edu

ISSN: 1044-3983/06/1703-0276
DOI: 10.1097/01.ede.0000208475.99429.2d

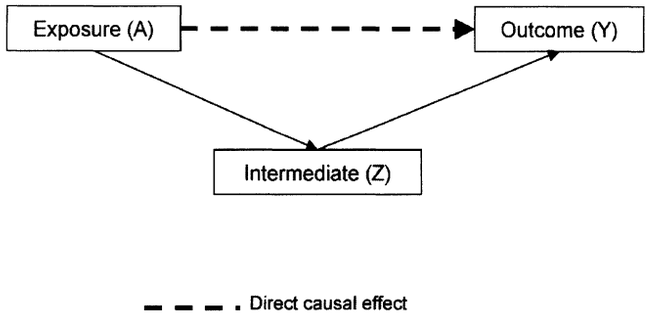

FIGURE 1. Basic causal structural of direct effect questions. Example 1: A = type of antiretroviral therapy (PI-based or not), Z = viral load, Y = CD4 T-cell count. Example 2: A = pollution level, Z = use of rescue medication, Y = lung function.

presented in an appendix). Estimation of both controlled and natural direct effects requires assumptions beyond those necessary to estimate total causal effects. We review these assumptions and present additional assumptions required for estimation of natural direct effects, which are less restrictive than corresponding assumptions considered in the current literature.[2–4] Finally, we review direct effect estimation in the context of a causal intermediate that is also a confounder.

NATURAL AND CONTROLLED DIRECT EFFECTS

Consideration of 2 alternative hypothetical experiments can help to clarify the distinction between controlled and natural direct effects. To estimate a controlled direct effect, a researcher would measure the effect of an exposure while holding the intermediate variable at a fixed level (Fig. 2A). In contrast, to estimate a natural direct effect, a researcher would measure the effect of an exposure, blocking the exposure's effect on the intermediate variable but allowing the intermediate to vary among individuals (Fig. 2B). These alternative direct effect definitions can be formalized using the counterfactual framework for causal inference.

Under the counterfactual framework, the causal effect of an exposure on an individual is defined as the difference in outcome if the same individual was exposed versus unexposed. These outcomes are termed counterfactual because only one outcome can be observed for a given individual; thus, causal inference can be viewed as a missing data problem. The counterfactual outcome under a given exposure $A = a$ is denoted Y_a, whereas Y_0 denotes a counterfactual outcome at the reference level of the exposure. In this article, we use "exposed" and "unexposed" as relative terms, in which "exposed" refers to the level of exposure that is of interest and "unexposed" ($A = 0$) refers to the reference level of exposure rather than implying the complete absence of exposure.

The counterfactual framework can also be used to define both types of direct causal effects. The controlled direct effect of an exposure on an individual is defined as the difference in counterfactual outcome if the individual was unexposed and her intermediate variable was controlled (or set) at level $Z = z$ versus the counterfactual outcome if she was exposed and her intermediate variable set at the same

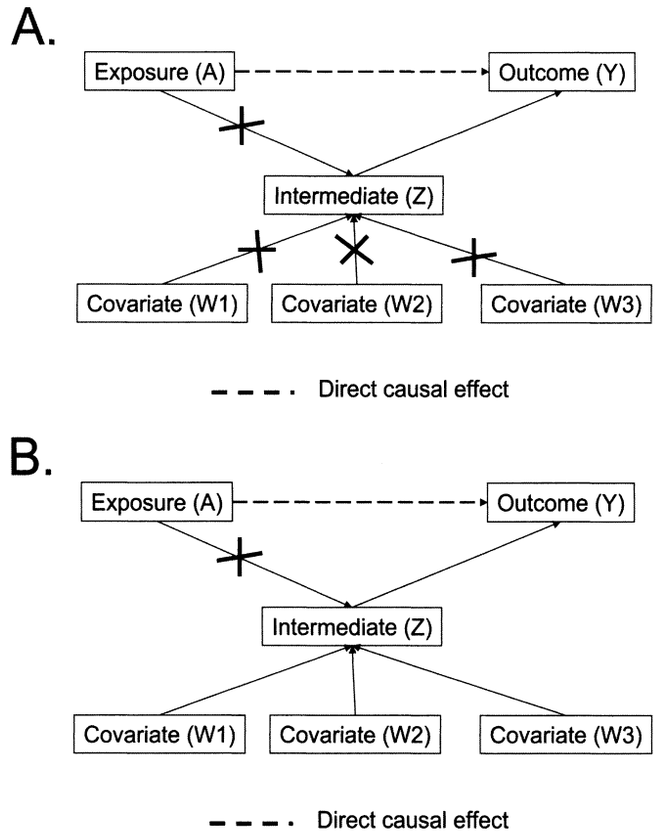

FIGURE 2. (**A**) Controlled direct effect of A on Y, holding Z at a fixed level (blocking all effects on Z). (**B**) Natural direct effect of A on Y, blocking only the effect of A on Z.

level $Z = z$. Using standard notation, the controlled direct effect of an exposure a on an individual is written $Y_{az} - Y_{0z}$, in which Y_{az} denotes an individual's counterfactual outcome controlling both the exposure and the intermediate variable. Note that these outcomes are counterfactual because neither the level of the exposure nor the level of the intermediate need correspond to the observed values of these variables.

Alternatively, the natural direct effect of an exposure on an individual is defined as the difference in counterfactual outcome if the individual was unexposed versus the counterfactual outcome if she was exposed, but her intermediate remained at its counterfactual level under no exposure.[2–4] To formally define a natural direct effect requires considering an additional counterfactual, the counterfactual level of an individual's intermediate at a given level of exposure $A = a$, denoted Z_a. The natural direct effect of an exposure a on an individual is written $Y_{aZ_0} - Y_{0Z_0}$, in which Z_0 is an individual's counterfactual level of the intermediate at the reference level of exposure.

Under both definitions, the population direct effect is the mean (or some other parameter) of the population distribution of the individual direct effects.

$$\text{Controlled direct effect: } E(Y_{az} - Y_{0z}) \qquad (1)$$

$$\text{Natural direct effect: } E(Y_{aZ_0} - Y_{0Z_0}) \quad (2)$$

Direct effects can be conceptualized as hypothetical experiments to make the definition of the corresponding counterfactuals more intuitive. Here, we apply this approach to our examples to help illustrate the 2 types of direct effect.

In example 1, the natural direct effect of protease inhibitor-based antiretroviral therapy (PI-based ART) on CD4 T-cell count is defined as the difference in average CD4 T-cell count if the study population was treated with PI-based versus non-PI-based ART, and viral load remained at the level it would have had on non-PI-based therapy. In contrast, the controlled direct effect is the effect of PI-based ART if viral load was controlled at a single level for all individuals.

The controlled direct effect corresponding to example 2 is the change in the expected lung function if the entire population were exposed to an incremental increase in air pollution and every member of the population was to use a single fixed level of rescue medication. The natural direct effect in this example is the change in expected lung function if the entire population were exposed to an incremental increase in air pollution but every individual in the population continued to use rescue medication as he would have had air pollution not increased.

Choice among estimation of the controlled direct effect, the natural direct effect, or both must depend on the research question. When an exposure and an intermediate variable interact to cause the outcome, estimation of the controlled direct effect depends on the level at which the intermediate variable is fixed, whereas the natural direct effect provides a single summary of the direct effect in the study population. For example, the direct effect of air pollution might differ at different levels of rescue medication use; the natural direct effect provides a population-level summary of the impact of air pollution (of course, there may be other interesting summaries as well).

Natural direct effects may be of particular interest in research contexts in which the researcher finds it helpful to conceive of the causal intermediate varying between individuals. For example, it may not be logical to think of fixing the rescue medication of the entire population at a given level. There are likely to be children in the population with underlying respiratory diseases who will always require rescue medication, regardless of air pollution levels; the direct effect of air pollution if these children (along with all others in the population) did not use medication may be difficult to interpret. Pearl discusses additional situations in which natural direct effects may be of particular interest.[3]

ESTIMATION OF DIRECT EFFECTS

Under the assumptions discussed in the following sections, and when exposure and intermediate do not interact to affect the outcome, standard multivariable regression of outcome on exposure, intermediate, and all confounders provides an estimate of both the controlled and natural direct effects. When exposure and intermediate do interact to affect outcome, such a regression estimates the controlled direct effect. Testing the null hypothesis that no controlled direct effect is present at any level of the intermediate (by testing whether the coefficients on all terms containing the exposure of interest in the multivariable model equal zero) also corresponds to a valid test of the null hypothesis that no natural direct effect is present. However, in settings where the exposure and intermediate do interact, additional steps are needed to estimate the natural direct effect.

In this section, we discuss estimation of natural direct effects using an example based on the effect of PI-based ART on CD4 T-cell count (example 1). For a formal presentation of the approach, as well as a numeric illustration, see the Appendix. The data consist of the following (Fig. 3):

- An outcome (Y), CD4 T-cell count;
- An intermediate (Z), viral load;
- A binary exposure (A), PI-based ART; and
- A confounder (W), an indicator of past treatment with mono/dual ART.

In the current example, we assume only a single confounder; however, the same methods and assumptions apply to contexts with multiple confounders.

Under assumptions discussed in the following sections, the natural direct effect is identified by the following formula[3,4]:

$$DE(a) = E_w \sum_z \{E(Y_{az} \mid W) - E(Y_{0z} \mid W)\} \Pr(Z_0 = z \mid W) \quad (3)$$

This formula says that the natural direct effect, among subjects who have identical values of all confounders W, can be calculated as a weighted average of the controlled direct effect at each possible level of the intermediate with the weight for a given level of the intermediate determined by the probability that the intermediate would have taken that level if the exposure were set at its reference level. Note that this probability can depend on the values of the confounders. This yields a separate estimate of the natural direct effect for each subgroup defined by the confounders. To estimate the natural direct effect in the whole population simply requires taking an additional weighted average of these subgroup-specific direct effects with the weight for a given subgroup determined by the relative size of the subgroup in comparison with the population.

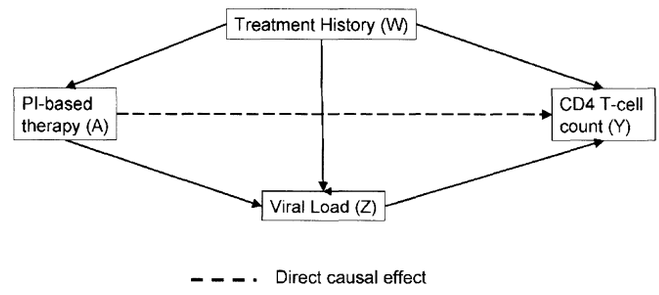

FIGURE 3. Causal structure for HIV example.

Under the assumptions discussed subsequently, the natural direct effect of an exposure can be estimated using standard statistical methods. Estimation of the natural direct effect begins with multivariable regression of outcome on intermediate, exposure and confounders. Regression of CD4 T-cell count on PI-based ART use, viral load, and history of mono/dual therapy gives an estimate of the direct effect of PI-based ART at a controlled level of viral load within subgroups defined by confounder values. An additional multivariable regression of the intermediate on exposure and confounders is then used to estimate the expected level of the intermediate variable at the reference level of exposure within subgroups defined by confounder values; regression of viral load on treatment history and use of PI-based ART allows us to estimate viral load under non-PI-based ART. The marginal direct effect of the exposure in the study population is estimated with the level of the intermediate controlled at its expected level at the reference level of exposure.

ASSUMPTIONS FOR THE ESTIMATION OF CONTROLLED AND NATURAL DIRECT EFFECTS

Estimation of both controlled and natural direct effects requires assumptions beyond those needed to estimate total causal effects. Like in any attempt to estimate a causal effect using multivariable regression, one must assume that there is no residual confounding of the effect of the exposure on the outcome beyond the covariates included in the model. As represented in the DAG framework, the standard assumption of no unmeasured confounders requires the absence of any unmeasured covariate that is a cause of both the exposure and outcome (*U*1 in Fig. 4).

However, consistent estimation of both controlled and natural direct effects also requires the additional assumption of no residual confounding of the effect of the intermediate on the outcome.[2,5,6] In other words, one must assume that, within subpopulations defined by regression covariates and exposure status, there are no unmeasured variables that predict the level of the intermediate variable and, independently, predict the outcome. In Figure 4, this assumption corresponds with the absence of any unmeasured covariate that is a cause of both the intermediate variable and the outcome (*U*2 in Fig. 4).

Formally, the assumptions regarding no unmeasured confounders can be summarized as follows (where $X \amalg Y$ means X is independent of Y):

$$A \amalg Y_{az} \mid W \tag{4}$$

$$Z \amalg Y_{az} \mid A,W \tag{5}$$

Several authors present techniques aimed at estimating controlled direct effects in the presence of unmeasured confounders.[7,8] In addition, see Rubin for a discussion of direct effects based on principal stratification.[9]

ADDITIONAL ASSUMPTIONS FOR THE ESTIMATION OF NATURAL DIRECT EFFECTS

In addition to the assumptions of no unmeasured confounders of either the effect of the exposure on the outcome or the effect of the intermediate on the outcome, estimation of natural direct effects requires 2 additional assumptions.

First, we assume that there are no unmeasured confounders of the effect of the exposure on the intermediate variable Z (*U*3 in Fig. 5A). This assumption is necessary to ensure that regressing the intermediate on the exposure and covariates and evaluating the resulting model with exposure set equal to its reference level is providing a consistent estimate of the counterfactual level of the intermediate variable at the reference level of exposure. For example, if a poor ability to adhere to prescribed medications results in both a

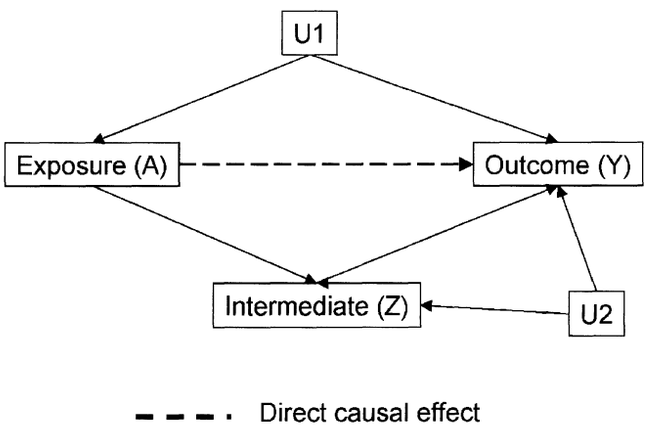

FIGURE 4. Unmeasured confounders of exposure effect (U1) and intermediate effect (U2).

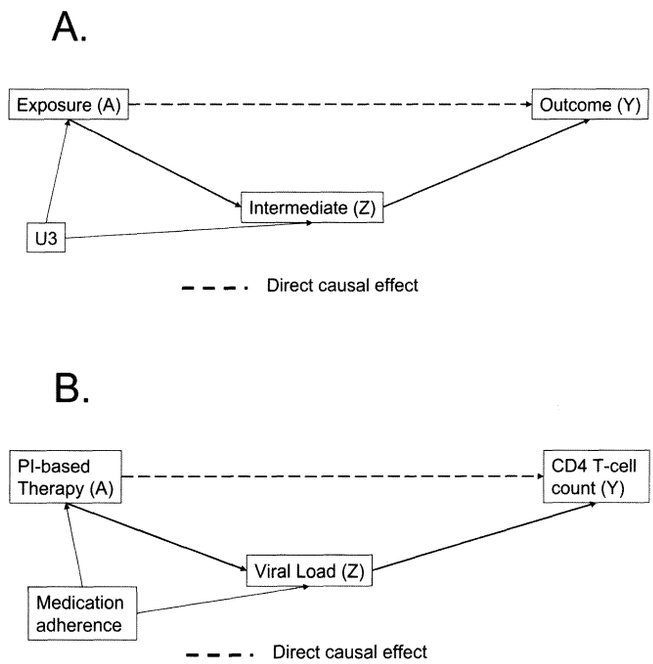

FIGURE 5. (**A**) Confounding of the effect of exposure on the intermediate variable. (**B**) Confounding of the effect of exposure on the intermediate variable. Illustration based on example 1.

higher viral load and an increased probability of assignment to a PI-based regimen, failure to include adherence when regressing viral load on regimen and treatment history will result in an underestimate of the counterfactual viral load that would have been observed if the entire study population had received a non-PI-based therapy (Fig. 5B). Formally, this additional assumption states:

$$A \amalg Z_a \mid W \tag{6}$$

Second, we assume that, within subgroups defined by covariates included in our multivariable model, the level of the intermediate variable in the absence of exposure does not tell us anything about the expected magnitude of the exposure's effect at a controlled level of the intermediate variable. We refer to this assumption as the direct effect assumption, which can be stated formally as:

$$E(Y_{az} - Y_{0z} \mid Z_0 = z, W) = E(Y_{az} - Y_{0z} \mid W) \tag{7}$$

In our example, the direct effect assumption states that, within strata defined by treatment history, knowing what an individual's viral load would have been on non-PI-based treatment does not provide any additional information about the effect of PI-based treatment on the individual's expected CD4 T-cell count at a controlled viral load. In any causal inference problem, we only observe a single outcome for a given individual, the counterfactual outcome corresponding to the treatment the individual actually received. Estimation of direct effects requires the additional identifying assumption (7) because one of the counterfactual outcomes, Y_{aZ_0}, corresponds to 2 different treatments in the same individual, and as a result is never observed.

Previous work discussing estimation of natural direct effects suggested that the assumptions necessary to make these effects identifiable were more restrictive than those presented here.[2–4] Robins and Greenland proposed an alternative to our direct effect assumption (7),[2,4] which states that the intermediate and the exposure of interest do not interact to affect outcome at the individual level. Such an assumption is both very restrictive and unrealistic in many biologic settings and can be tested by examining the data. In our example, evidence supporting heterogeneity of the effect of PI-based ART depending on the level of viral load would suggest that the assumption is violated. Note that if the "no interaction" assumption were to hold, estimation of both controlled and natural direct effects would simply require taking a weighted average of the confounder-specific direct effect estimates generated by a single regression of outcome on exposure, intermediate, and confounders.

Pearl proposed a third alternative identifying assumption,[3] which states that, within subgroups defined by baseline covariates included in the model, an individual's counterfactual outcome does not depend on the level of the intermediate in the absence of exposure:

$$Y_{az} \amalg Z_0 \mid W \tag{8}$$

An alternative way of formulating this assumption is that, within subgroups defined by baseline covariates, individual counterfactual outcome is a deterministic function of treatment, the level of the intermediate, and an exogenous error (conditionally independent of Z_0 given W), but not of the counterfactual outcome under no treatment. In contrast, under our assumption, at a controlled level of z, an individual's counterfactual outcome under a given treatment, Y_{az}, can depend on the individual's counterfactual outcome under no treatment, Y_{0z}. Generally, Y_{0z} explains a lot of the variation in Y_{az}, suggesting that our assumption is more reasonable. In addition, it can be shown that assumption (7) holds in essentially all cases assumption (8) holds and in many cases where it does not.

We refer readers to the Appendix for an in-depth comparison of our assumptions with the assumptions of previous authors. In conclusion, we note that, even when our direct effect assumption (7) fails to hold, equation (3) still estimates an interesting causal parameter: a summary of the direct effect of the exposure in the population with the intermediate controlled at its mean counterfactual level in the absence of exposure.

CONFOUNDING BY A CAUSAL INTERMEDIATE

Assumptions (4) and (5) are sufficient to ensure the identifiability of controlled direct effects and, in combination with assumptions (6) and (7), are sufficient to ensure the identifiability of natural direct effects. However, these assumptions alone do not imply that the standard multivariable regression techniques presented are necessarily adequate to estimate either type of direct effect. As discussed by Robins and Greenland, in the case in which a confounder of the effect of the intermediate variable is itself affected by the exposure of interest, traditional multivariable methods will provide a biased estimate of the controlled direct effect.[2] A confounder of this type is illustrated in Figure 6; a variable ("C") affects both the intermediate variable and the outcome of interest, and so acts as a confounder of the effect of the intermediate variable; in addition, the confounder "C" is itself a causal intermediate between the exposure and intermediate variable.

The analytic dilemma posed by confounding of a direct effect in a single time point study by a variable that is itself affected by the exposure of interest is similar to the problem of time-dependent confounding that frequently occurs when estimating total effects in a longitudinal data setting.[10] In general, if there is a variable that affects both the intermediate variable and

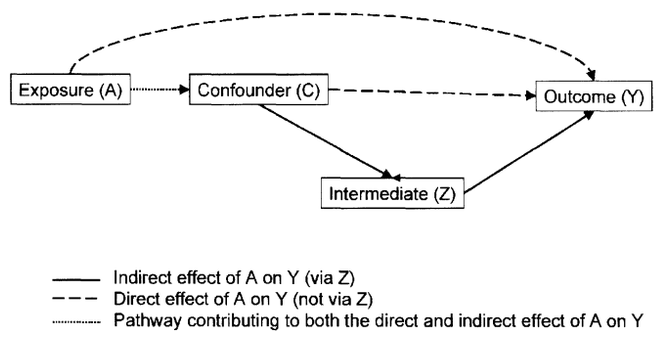

FIGURE 6. Confounding by a causal intermediate.

the outcome, the effect of the intermediate on the outcome will be confounded unless this variable is included in the multivariable regression model. However, if this confounding variable is itself affected by the exposure, including it in the multivariable regression model will also result in a biased estimate of effect by controlling for part of the effect of interest. Consistent estimation of controlled direct effects in the context of confounding by a causal intermediate requires nonstandard techniques such as g-computation or inverse probability weighting.[2,11] The resulting estimates can then be used to estimate natural direct effects according to formula (3).

DISCUSSION

The estimation of direct effects is a common goal in epidemiologic research. In the estimation of direct effects, like in all analyses, the choice of method must be driven by the research question. In settings where the aim is to estimate the causal effect of an exposure while holding the level of the intermediate variable at a controlled level defined by the researcher (controlled direct effect), multivariable regression, under assumptions, may indeed provide an estimate of the effect of interest. However, if the research goal is to estimate the effect of an exposure on an outcome if the exposure's effect on the intermediate was blocked, allowing the intermediate to follow the course it would have taken at the reference level of exposure (natural direct effect), then multivariable regression alone may be insufficient. In the case in which exposure and intermediate do not interact at the individual level to affect outcome, controlled and natural direct effects are equivalent; however, such an assumption is not required to ensure their identifiability.

Previous literature has suggested that estimation of direct effects is either impractical given the assumptions required or uninteresting given the lack of correspondence to real-world problems. We have proposed a less restrictive identifying assumption and explained interpretation of the direct effect parameter even when this assumption is not met. We further argue that lack of a corresponding real-world experiment need not imply that a causal parameter is uninteresting. In the examples discussed, neither type of direct effect corresponds to a realistic hypothetical experiment. However, this does not imply that estimation of these direct effects is not of interest. The goal in these examples is to understand a drug mechanism (example 1) and to understand the impact that pollution has on children's health (example 2).

Rather than the plausibility of a corresponding experiment, the central issue is whether or not the researcher finds the counterfactual definitions of direct effects interpretable. Many epidemiologic examples exist in which the counterfactual, or missing data, model is widely accepted despite the absence of any corresponding real-world experiment. (For example, survival analysis assumes that an individual has an underlying survival time that could have been observed if he had not been censored by, say, being hit by a truck. Most researchers are perfectly happy with this assumption, despite the absence of any intervention that could prevent all possible deaths by truck collision.) It is up to the researcher to determine whether the counterfactual model is interpretable for a given application.

Even in the case in which the researcher is uncomfortable with a counterfactual interpretation of the direct effect parameters, the direct effect parameters can still be interesting and interpretable. In this case, under the assumptions of no unmeasured confounders alone (4, 5, and 6), a controlled direct effect can be understood as the difference in outcome between individuals who are exposed versus unexposed and exchangeable with respect to (having the same values of) all confounders and the level of the intermediate variable. Similarly, according to equation (3), the natural direct effect is simply an average of the controlled direct effect at each level of the intermediate weighted with respect to the distribution of the intermediate variable in the unexposed. In the estimation of direct (as well as total) causal effects, the counterfactual framework can be viewed as simply a helpful tool for identifying interesting parameters of the data generating distribution; ultimately, it is the researcher who must decide the suitability of a counterfactual interpretation of these parameters.

In summary, we feel that the barriers to estimation of direct effects are not as great as has been suggested. We encourage researchers to estimate direct effects of interest while giving appropriate consideration to the relevant assumptions and the resulting interpretation of their estimates.

ACKNOWLEDGMENTS

We thank James Robins for insightful discussion.

REFERENCES

1. Greenland S, Pearl J, Robins JM. Causal diagrams for epidemiologic research. *Epidemiology*. 1999;10:37–81.
2. Robins JM, Greenland S. Identifiability and exchangeability for direct and indirect effects. *Epidemiology*. 1992;3:143–155.
3. Pearl J. Direct and indirect effects. In: *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*. San Francisco: Morgan Kaufmann; 2001:411–420.
4. Robins JM. Semantics of causal DAG models and the identification of direct and indirect effects. In: Hjort N, Green P, Richardson S, eds. *Highly Structured Stochastic Systems*. Oxford: Oxford University Press; 2003:70–81.
5. Cole SR, Hernan MA. Fallibility in estimating direct effects. *Epidemiology*. 2002;31:163–165.
6. Poole C, Kaufman JS. What does standard adjustment for down- stream mediators tell us about social effect pathways. *Am J Epidemiol*. 2000; 151:s52.
7. Joffe MM, Colditz GA. Restriction as a method for reducing bias in the estimation of direct effects. *Stat Med*. 1998;17:2233–2249.
8. Kaufman S, Kaufman JS, Maclehose RF, et al. Improved estimation of controlled direct effects in the presence of unmeasured confounding of intermediate variables. *Stat Med*. 2005;24:1683–1702.
9. Rubin DB. Direct and indirect causal effects via potential outcomes. *Scandinavian Journal of Statistics*. 2004;31:161–170.
10. Robins JM, Hernan MA, Brumback B. Marginal structural models and causal inference in epidemiology. *Epidemiology*. 2000;11:550–560.
11. van der Laan MJ, Petersen ML. Estimation of direct and indirect causal effects in longitudinal studies. Technical report, University of California, Berkeley, Division of Biostatistics. August 23, 2004. Available at: http://www.bepress.com/ucbbiostat/paper155.
12. Phenix B, Angel J, Mandy F, et al. Decreased HIV-associated T-cell apoptosis by HIV protease inhibitors. *AIDS Res Hum Retroviruses*. 2000;16:559–567.
13. Muthumani K, Choo A, Hwang D, et al. Mechanism of HIV-1 viral protein R-induced apoptosis. *Biochem Biophys Res Commun*. 2003;304: 583–592.

APPENDIX

Numeric Example Illustrating Implementation of the Natural Direct Effect Estimate

Implementation of the direct effect estimate involves the following steps:

1. Fit a multivariable regression of outcome on confounders, exposure, and intermediate variable. For example, we fit the following linear regression model of CD4 T-cell count (Y) on viral load (Z), treatment history (W), and PI-based ART (A).

$$\hat{E}(Y \mid A,W,Z) = 450 + 50A - 20AW + 10AZ + 10AZW - 50W - 100Z \quad (9)$$

2. Estimate the direct effect of the exposure (given W) at a controlled level of the intermediate variable (controlled direct effect): $E(Y_{az} - Y_{0z} \mid W)$. In our example,

$$\hat{E}(Y_{1z} - Y_{0z} \mid W) = \hat{E}(Y \mid A = 1,Z = z,W) - \hat{E}(Y \mid A = 0,Z = z,W) = 50 - 20W + 10z + 10zW \quad (10)$$

3. Estimate an individual's natural direct effect by replacing z in equation (10) with an estimate of the level of the intermediate variable that the individual would have had in the absence of exposure (Z_0) and estimate the population direct effect as the mean of the individual direct effects.

$$\hat{E}(Y_{1Z_0} - Y_{0Z_0}) = 50 - 20\hat{E}(W) + 10\hat{E}(Z_0) + 10\hat{E}(Z_0W) \quad (11)$$

4. Estimate $E(Z_0)$ by fitting a multivariable regression of the intermediate on exposure and confounders. In our example, we regress viral load on treatment history and PI-based ART use.

$$\hat{E}(Z \mid A,W) = 1.7 + 1.25W + 0.2A + 0.2AW \quad (12)$$

The average of $E(Z \mid A = 0,W)$ across the population provides an estimate of $E(Z_0)$. In our example, 33% of the study population have a history of mono/dual therapy ($W = 1$).

$$\hat{E}(W) = \Pr(W = 1) = 0.33 \quad (13)$$

Thus, the average predicted viral load in the study population under non-PI-based ART is

$$\hat{E}(Z_0) = \hat{E}(\hat{E}(Z \mid A = 0,W)) = \hat{E}(Z \mid A = 0,W = 1)\hat{\Pr}(W = 1) + \hat{E}(Z \mid A = 0,W = 0)\hat{\Pr}(W = 0) = 2.95 \times 0.33 + 1.7 \times 0.67 = 2.1$$

5. Similarly, the average of $E(Z \mid A = 0,W) \times W$ across the population provides an estimate of $E(Z_0W)$. In our example,

$$\hat{E}(Z_0W) = \hat{E}(\hat{E}(Z \mid A = 0,W) \times W) = \hat{E}(Z \mid A = 0,W = 1) \times (1) \times \hat{Pr}(W = 1) + \hat{E}(Z \mid A = 0,W = 0) \times (0) \times \hat{\Pr}(W = 0) = 2.95 \times 0.33 = 0.97$$

6. Substitute these values into model (11) to get an estimate of the natural direct effect in the population.

$$\hat{DE} = \hat{E}(Y_{1Z_0} - Y_{0Z_0}) = 50 - 20\hat{E}(W) + 10\hat{E}(Z_0) + 10\hat{E}(Z_0W) = 50 - 20 \times 0.33 + 10 \times 2.1 + 10 \times 0.97 = 74.1$$

In our example, we estimate that PI-based ART has a natural direct effect of 74 CD4 T-cells.

Estimation Approach for Natural Direct Effects

Consider the simple single time point data structure W, A, Z, Y. For simplicity, we assume that all variables are univariate. We assume that within strata of W, (A,Z) is randomized (there is no unmeasured confounding at the level of either treatment or the intermediate variable), and our direct effect assumption (7). In the single time point case, given that no confounders are also affected by the exposure of interest, one can then use standard regression methods to test for and estimate a natural direct effect.

Under the assumption that (A,Z) is randomized with respect to W, we have $E(Y \mid A = a,Z = z,W) = E(Y_{az} \mid W)$ and thus that $E(Y_{az} - Y_{0z} \mid W) = E(Y \mid A = a,Z = z,W) - E(Y \mid A = 0,Z = z,W)$. We now assume a linear regression model for $E(Y \mid A, Z, W) = A(\beta_0 + \beta_1Z + \beta_2W + \beta_3ZW) + (\alpha_0 + \alpha_1Z + \alpha_2W + \alpha_3ZW)$, so that we have the model

$$E(Y_{az} - Y_{0z} \mid W) = a(\beta_0 + \beta_1z + \beta_2W + \beta_3zW). \quad (14)$$

One can then test for no direct effect by testing H_0: $\beta_0 = \beta_1 = \beta_2 = \beta_3 = 0$.

To estimate the natural direct effect, we need to take the conditional expectation of (14) over z with respect to the distribution of Z_0, given W. For this purpose, we assume a model $m(a,W \mid \lambda)$ for $E(Z_a \mid W) = E(Z \mid A = a,W)$ indexed by parameters λ. By the linearity of (14) in z, it follows that the direct effect is now modeled as $DE = A(\beta_0 + \beta_1E(m(0,W \mid \lambda) + \beta_2EW + \beta_3E(m(0,W \mid \lambda)W))$, where $m(0,W \mid \lambda)$ is the model of the expected value of the counterfactual intermediate variable, given treatment and baseline covariates, evaluated at $a = 0$. An estimate of DE is obtained by replacing the regression parameters (β,λ) by their least squares estimators.

Comparison With Identifying Assumptions in Prior Literature

We propose the following novel assumption for identification of natural direct causal effects:

$$E(Y_{az} - Y_{0z} \mid Z_0 = z,W) = E(Y_{az} - Y_{0z} \mid W), \quad (7)$$

for all a and z.

Comparison With Pearl[3]

Pearl shows (using the structural equation framework) that the natural direct effect is identifiable, if

$$Y_{az} \amalg Z_0 \mid W \quad (8)$$

for all z.

It is of interest to compare our assumption (7) with Pearl's assumption (8).

An alternative way of formulating assumption (8) is that there exists a function m such that

$$Y_{az} = m(a,z,W,e), \quad (15)$$

where e is a random variable, which is conditionally independent of Z_0, given W. Stated in words, Pearl assumes that, within subgroups defined by baseline covariates, individual counterfactual outcome is a deterministic function of treatment, the level of the intermediate variable, and an exogenous error, but not of the counterfactual outcome under the reference treatment. In contrast, under our assumption, at a fixed level of z, an individual's counterfactual outcome under a given treatment, Y_{az}, can depend on the individual's counterfactual outcome under the reference treatment, Y_{0z}. Generally, Y_{0z} explains a lot of the variation in Y_{az}, suggesting that our assumption is more reasonable.

For example, subjects can have different CD4 T-cell counts under non-PI-based therapy (Y_{0z}), which are themselves extremely predictive of the counterfactual CD4 T-cell count Y_{az} under PI-based therapy and are not explained by baseline covariates W. In other words, within subpopulations defined by baseline covariates W and a fixed viral load z, an individual's CD4 T-cell count on PI-based therapy is likely to depend on what that individual's CD4 T-cell count would have been under non-PI-based therapy. In this case, the assumption of Pearl does not hold. However, it seems less unreasonable to assume that, within subpopulations defined by baseline covariates and fixed viral load z, the average magnitude of the controlled direct effect of PI-based antiretroviral therapy does not differ between individuals with different CD4 T-cell counts under non-PI-based therapy.

Suppose that assumption (8) holds at 2 treatment values a and 0. In that case, we have that both counterfactual outcomes Y_{az} and Y_{0z} are conditionally independent of Z_0, given W. One would now expect that the difference $Y_{az} - Y_{0z}$ is also conditionally independent of Z_0, given W, and thus for our assumption (7) to hold. (In fact, mathematically it follows that $Y_{az} - Y_{0z}$ is uncorrelated with any real valued function of Z_0, given W.) This suggests that in most examples in which assumption (8) holds, our assumption (7) also holds. On the other hand, it is easy to construct examples in which our assumption holds, whereas assumption (8) fails to hold.[11] We refer to Robins for further discussion of the limitations of assumption (8).[4]

Comparison With Robins[2]

Robins and Greenland propose an alternative identifying assumption, which they call the no-interaction assumption:

$$Y_{az} - Y_{0z} \text{ is a random function } B(a) \text{ that does not depend on } z \quad (16)$$

In words, this assumption states that the individual controlled direct effect at a fixed-level z does not depend on the level at which z is fixed, or in other words, that the intermediate variable does not interact with the exposure of interest in its effects on outcome.

A detailed mechanistic discussion of this assumption is given in Robins and Greenland.[2] The "no-interaction assumption" implies, in particular, that $E(Y_{az}) = m_1(a) + m_2(z)$ for some functions m_1 and m_2, or in other words, that the marginal causal effects of the treatment and the intermediate variable on outcome are additive. In most applications, one expects these interactions to be present, and, the interactions themselves often correspond with interesting statistical hypotheses. Consequently, the "no-interaction assumption" is very restrictive.

Applied to our HIV example, assumption (16) implies that the individual direct effect of PI-based antiretroviral treatment at a controlled viral load does not depend on the level at which viral load is controlled. In other words, the direct effect of PI-based treatment on CD4 T-cell count would be the same if viral load was controlled at a high level or controlled at a low level. This assumption is unlikely to be met and is an interesting research question in itself. In particular, PI-based regimens are hypothesized to act directly on CD4 T-cells by inhibiting their apoptosis (programmed cell death).[12] Higher levels of ongoing CD4 T-cell apoptosis may be induced by higher viral loads.[13] Thus, we would expect that if PI-based therapy has an antiapoptotic direct effect on CD4 T-cell count (ie, not mediated by changes in viral load), such an effect might be larger among individuals with higher viral loads and higher levels of apoptosis. In such a case, Robins' assumption does not hold.

Identifiability Result

Under the direct effect assumption (7), we have the following identifiability result for natural direct effects, presented as a theorem:

Theorem 1: *Let* $DE(a) = E(Y_{aZ_0} - Y_{0Z_0})$. *Assume the assumption (7) holds. Then*

$$DE(a) = \tilde{D}E(a) \equiv E_W\int\{E(Y_{az} \mid W) - E(Y_{0z} \mid W)\}dF_{Z_0|W}(z)$$

Proof:

$$\begin{aligned} DE(a) = E(Y_{aZ_0} - Y_{0Z_0}) &= E_W(E(Y_{aZ_0} - Y_{0Z_0} \mid W)) \\ &= E_W(E_{Z_0|W}(E(Y_{aZ_0} - Y_{0Z_0} \mid Z_0,W))) \\ &= E_W\int_z\{E(Y_{az} - Y_{0z} \mid Z_0 = z,W)\}df_{Z_0|W}(z) \\ &\overset{(7)}{=} E_W\int_z\{E(Y_{az} - Y_{0z} \mid W)\}dF_{Z_0|W}(z) = \tilde{D}E \quad (V) \end{aligned}$$

where the righthand side is identifiable from the observed data distribution.

The identifiability result of Pearl can be shown in precisely the same manner. Clearly, the assumption of Pearl (8) also implies $DE = \tilde{DE}$. Thus, Pearl's identifiability mapping is the same as ours (theorem 1), but it was based on a different assumption. Similarly, under the assumption of Robins (16), we have $Y_{aZ_0} - Y_{0Z_0} = Y_{az} - Y_{0z}$ for any z so that

$$E(Y_{aZ_0} - Y_{0Z_0}) = E(Y_{az} - Y_{0z}) \quad (17)$$

where the latter quantity does not depend on z. Robins' identifiability mapping (17) corresponds with ours using an empty W (and thus with Pearl's), because the integration with respect to F_{Z_0} does not affect the integral. We conclude that all 3 identifiability mappings agree with each other (except that Robins avoids integration with respect to F_{Z_0} by making the "no-interaction assumption" [16]), but that the model assumptions that were used to validate the identifiability mapping are different. Our result shows that the identifiability mapping of Pearl holds under a less restrictive *union assumption*: that is, the identifiability result presented in theorem 1 holds if either our assumption (7) holds, or the assumption (8) holds, or the "no-interaction assumption" (16) holds.

In the Next Issue

Perceived stress and risk of ischemic heart disease: Causation or bias?

Reliability and validity of 2 single-item measures of psychosocial stress

Will low participation in cohort studies induce bias in relative risks?

Feasibility of the current-duration approach to studying human fecundity

Validity issues relating to time-to-pregnancy studies of fertility

Semen quality and persistent organochlorine pollutants in an Inuit and 3 European cohorts

PBBs, PCBs, body weight and incidence of adult-onset diabetes mellitus

Association of nitrate and nitrite from drinking water and diet with risk of non-Hodgkin lymphoma

Modification of risk of arsenic-induced skin lesions by sunlight exposure, smoking and occupational exposures in Bangladesh

Intra-individual variability of plasma antioxidants, markers of oxidative stress, C-reactive protein, cotinine and other biomarkers

Concordance rates and modifiable risk factors for lower urinary tract symptoms in twins

Aspirin use and miscarriage

Adult weight change, weight cycling and pre-pregnancy obesity in relation to risk of preclampsia