Estimation of Direct Causal Effects

Maya L. Petersen[*] Mark J. van der Laan[†]

[*]Division of Biostatistics, School of Public Health, University of California, Berkeley, mayaliv@berkeley.edu

[†]Division of Biostatistics, School of Public Health, University of California, Berkeley, laan@berkeley.edu

Estimation of Direct Causal Effects

Maya L. Petersen 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 arises frequently in 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, multivariable regression estimates a particular type of direct effect, the effect of an exposure on outcome fixing the intermediate at a specified level. Using the counterfactual framework, we distinguish this definition of a direct effect (Type 1 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 (Type 2 direct effect). When the intermediate and exposure interact to affect the outcome these two types of direct effects address distinct research questions. Relying on examples, we illustrate the difference between Type 1 and Type 2 direct effects. We propose an estimation approach for Type 2 direct effects that can be implemented using standard statistical software and illustrate its implementation using a numerical example. We also review the assumptions underlying our approach, which are less restrictive than those proposed by previous authors.

1 Introduction.

Many research questions in epidemiology are concerned with understanding the causal pathways by which an exposure or treatment affects an outcome. Consider the following examples:

- Researchers often aim to understand the biological mechanisms by which a treatment slows disease progression or an exposure acts to cause or prevent disease.

  Example 1: In HIV-infected individuals, antiretroviral therapy preserves CD4 T-cell counts. Are these beneficial effects due entirely to reductions in plasma HIV RNA level (viral load)?

- Individuals and their physicians often alter their treatment decisions as a result of exposures. In such cases, researchers may be interested in estimating the causal effects of an exposure if the exposure's effect on treatment decisions were blocked.

  Example 2: Air pollution regulations rely, in part, on estimates of the effect of exposure to pollutant levels on lung function in children. However, high pollutant levels may cause children to increase their use of rescue medication. How would pollutant levels affect lung function if medication use were to remain at the same frequency in the population that it would have had in the absence of elevated pollutants?

- Surrogate markers of an outcome of interest are frequently used when the outcome itself is rare or expensive to measure. The quality of a surrogate marker can be assessed by estimating the extent to which the effect of an exposure on an outcome is captured by the exposure's effects on the surrogate.

  Example 3: In studying the effect of hormone therapy on risk of cardiovascular events, C-reactive protein has been suggested as a promising surrogate outcome.[1] To what extent does therapy affect the true outcome of interest, risk of cardiovascular event, via a pathway that does not involve C-reactive protein?

The above are just a few examples of a common causal structure underlying epidemiological problems, represented in Directed Acyclic Graph (DAG) form in Figure 1.[2] In the applications considered in this article, the exposure of interest acts on the outcome via two pathways, one in which the exposure affects an intermediate variable which in turn affects the outcome, and one in which the effects of the exposure do not occur via changes in the intermediate. In the above examples, the goal is to estimate the effect of the exposure on the outcome if its effect on the intermediate variable were blocked. Effects of this kind are referred to as direct effects.

Epidemiologists and others have generally used standard analytic approaches, such as multivariable regression, to estimate direct effects. In the single time-point case such approaches may 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, while raised previously,[3−6] may not be widely appreciated. In cases where the necessary assumptions are met, multivariable regression provides an estimate of the direct effect of an exposure, *at a fixed level of the intermediate variable.* 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.*[3] Note that, in the former definition, all causal effects on the intermediate are blocked, while in the latter case, only the effect of the exposure on the intermediate is blocked (Figures 2, 3).

The two definitions of a direct effect address different research questions, particularly when the effect of the exposure of interest is modified by the level of the intermediate variable. In this paper we use the examples above to illustrate the analytic issues surrounding the estimation and interpretation of direct effects in the single time-point case. We introduce a simple method for the estimation of the direct effect of an exposure on an outcome, blocking only the effect of the exposure on the intermediate variable, and discuss when such an approach might be preferable to standard multivariable regression. We also discuss the assumptions necessary for our estimation approach, which are less restrictive than the assumptions considered in the current literature.[3,4,7]

2 Unmeasured confounding of direct effects.

The standard approach to the estimation of direct effects in epidemiology involves multivariable regression of the outcome on the exposure of interest, confounders, and the causal intermediate. Even in settings where the exposure occurs at a single time-point, this approach requires assumptions beyond those needed to estimate total causal effects. As 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. Represented using 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 Figure 4).

However, consistent estimation of direct effects also requires the additional assumption of no residual confounding of the effect of the intermediate on the outcome.[3,5,6] In other words, one must assume that, within subpopulations defined by regression covariates and exposure status, there are no unmeasured variables that predict both 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 Figure 4).

3 Confounding by a causal intermediate.

When a direct effect estimate is biased as a result of confounding of the effect of the intermediate variable, simply including the additional confounders as covariates in a multivariable regression model may be sufficient to remove the bias (provided, of course, that the additional confounders are measured). However, as discussed by Robins and Greenland, in the case where 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 direct effect.[3] A confounder of this type is illustrated in Figure 5; 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.[8] We refer the interested reader to Robins and Greenland for a simulated data example of the bias that can result from traditional multivariable regression in this setting.[3] Here, we provide intuitive understanding of confounding of direct effects by a causal intermediate by considering Example 1.

In this example, the research aim is to investigate whether protease inhibitor (PI)- based antiretroviral therapy for HIV-infected patients has a direct effect on CD4 T-cells that is not mediated by changes in patients' viral loads. We elaborate the causal structure presented in Figure 1 to include the presence of viral mutations associated with antiretroviral resistance (Figure 6). Specifically, PI-based treatment results in resistance mutations which lower viral fitness.[9] As a result, these mutations may reduce viral load.[10] In addition, PI resistance mutations may also act to preserve CD4 T-cells by a pathway unrelated to changes in viral load.[11,12] As is clear from Figure 4, resistance mutations are a confounder of the effect of viral load on CD4 T-cell count; thus, failure to include mutations in a multivariable regression model will result in residual confounding and hence a biased estimate of the direct effect, as discussed in Section 2. However, including resistance mutations in the multivariable model will also result in a biased estimate of the direct effect by removing part of the causal effect of interest; by including resistance in the model, we effectively "set" or fix its level, and as a result lose any component of the direct effect of treatment that is mediated by changes in resistance mutations.

In general, if a variable exists that affects both the intermediate and 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 of interest, including it in the multivariable regression model can also result in a biased estimate of effect.

4 Interpretation of multivariable regression.

In the single time-point setting, under the assumptions of no unmeasured confounding at either the level of the exposure or the intermediate variable, and given that no confounder of the effect of the intermediate variable is itself a causal intermediate, then standard multivariable regression of outcome on exposure, intermediate, and all confounders provides a valid test of the null hypothesis of "no direct effect" (by testing whether the coefficients on all terms containing the exposure of interest in the multivariable model equal zero). Multivariable regression in this setting also provides an estimate of the effect of an exposure on an outcome, holding the level of the intermediate variable fixed at a given level. Depending on the research question, the direct effect of an exposure at a fixed level of the intermediate may or may not be the quantity of interest. Example 2 illustrates a setting in which the standard multivariable regression approach may not estimate the direct effect of interest.

4.1 Example 2: The direct effect of air pollution on childhood lung function.

In Example 2, the goal is to quantify the impact that air pollution has on children's lung function in a given population. Because air pollution can cause children to increase their use of rescue medication, which in turn can improve lung function and obscure the impact of air pollution, the effect of interest is the change in lung function that would result from an increase in air pollution if the use of rescue medication did not change.

Suppose that multivariable regression of lung function (Y) on air pollution level (A) and use of rescue medication (Z) yields the following fit. Assume no confounders (Figure 1).

$$E[Y|A,Z] = 50 - 15A + 3AZ + 12Z \tag{1}$$

Given that Model (1) is correctly specified, the change in expected lung function resulting from an incremental increase in air pollution at a fixed level of rescue medication use can then be estimated as

$$E[Y|A = a+1, Z] - E[Y|A = a, Z] = -15 + 3Z \tag{2}$$

Such a model fit suggests that the direct effect of air pollution depends on the frequency of rescue medication use in the population; if the entire population were to use rescue medication ($Z = 1$), air pollution would decrease lung function less than if the entire population did not use rescue medication ($Z = 0$). However, while this may be an interesting finding, it does not answer the research question of interest: How would an increase in air pollution affect lung function in the population if rescue medication use remained the same? In addition, it may not be logical to think of fixing the rescue medication of the entire population at a given level. For example, 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 is in this case not a meaningful quantity.

5 A formal definition of direct effects.

Consider two alternative ideal experiments a researcher might conduct to estimate a direct effect: 1) the researcher might measure the effect of an exposure while somehow holding the intermediate variable at a fixed level (Figure 2); or 2) the researcher might measure the effect of an exposure, blocking the exposure's effect on the intermediate variable, but allowing the intermediate to vary between individuals (Figure 3). Under assumptions, standard multivariable regression can be used to reproduce the results of experiment 1; however, when experiment 2 answers the scientific question of interest, standard multivariable regression is often insufficient. In order to clarify the difference between the two types of direct effects, we define them 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 were exposed vs. unexposed. These outcomes are termed counterfactual because only one is observed for a given individual. Typically, a counterfactual outcome under a given exposure $A = a$ is denoted Y_a. For example, for a binary exposure, Y_0 would denote an individual's outcome in the absence of the exposure and Y_1 the same individual's outcome in its presence.

The counterfactual framework can also be used to define both types of direct causal effects. In experiment 1, the direct effect of an exposure on an individual is defined as the difference in counterfactual outcome if the individual were exposed and her intermediate variable fixed at level $Z = z$ vs. the counterfactual outcome if she were unexposed and her intermediate fixed at the same level $Z = z$. Using standard notation, the direct effect of an exposure a on an individual can thus be written $Y_{az} - Y_{0z}$, where Y_{az} denotes an individual's counterfactual outcome controlling both exposure and intermediate variable. We refer to this definition as an individual Type 1 direct effect.

Alternatively, in experiment 2 the direct effect of an exposure on an individual is defined as the difference in counterfactual outcome if the individual were unexposed vs. the counterfactual outcome if she were exposed, but her intermediate variable remained at its counterfactual level under no exposure.[3,4,7] We refer to this definition as a Type 2 direct effect. To formally define a Type 2 direct effect requires considering an additional counterfactual, the counterfactual level of an individual's intermediate variable at a given level of exposure $A = a$, denoted Z_a. The Type 2 direct effect of an exposure a on an individual can be written $Y_{aZ_0} - Y_{0,Z_0}$, where Z_0 is an individual's counterfactual level of the intermediate in the absence 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.

$$\text{Type 1 Direct Effect: } E(Y_{az} - Y_{0z}) \tag{3}$$
$$\text{Type 2 Direct Effect: } E(Y_{aZ_0} - Y_{0,Z_0}) \tag{4}$$

To illustrate, in Example 2 the Type 1 direct effect 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 were forced to use the same fixed level of rescue medication. The Type 2 direct effect also estimates the change in expected lung function if the entire population were exposed to an incremental increase in air pollution, but allows every individual in the population to continue to use rescue medication as he did at the reference level of air pollution.

6 Estimation of Type 2 direct effects.

In the single time-point case (under assumption 14, see below), the Type 2 direct effect is identified by the following formula:

$$\text{DE}(a) \;=\; E_W \sum_z \{E(Y_{az} \mid W) - E(Y_{0z} \mid W)\}\, Pr(Z_0 = z|W). \tag{5}$$

In the case of a confounder that is affected by the exposure (Section 3), (5) still permits consistent estimation of Type 2 direct effects by assuming a marginal structural model and employing a corresponding estimation procedure (such as inverse probability weighting or g-computation) to estimate the quantity $E(Y_{az}|W) - E(Y_{0z}|W)$.[3,13] In the absence of a confounder that is also a causal intermediate, and under the assumptions of no unmeasured confounding, the Type 2 direct effect of an exposure can be estimated using standard statistical methods.

Estimation of the Type 2 direct effect begins with standard multivariable regression of outcome on intermediate, exposure and confounders. It then involves fitting an additional multivariable model regressing the intermediate on exposure and confounders. This latter model is used to predict each individual's expected level of the intermediate variable in the absence of exposure, based on that individual's covariates. The marginal direct effect of the exposure in the study population is estimated with the level of the intermediate fixed at its expected level in the absence of exposure.

We demonstrate how to implement this estimate using a numerical example based on the estimation of the direct effect of PI-based therapy on CD4 T-cell count (Example 1). (For a formal presentation of the approach, please see the online appendix). The data consist of an outcome, CD4 T-cell count, an intermediate, viral load, an exposure, PI-based therapy, and a confounder, an indicator of treatment with mono/dual antiretroviral therapy prior to baseline (Figure 7).

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 therapy (A).

   $$\hat{E}(Y|A,W,Z) = 450 + 50A - 20AW + 10AZ + 100AZW - 50W - 100Z \tag{6}$$

2. Estimate the direct effect of the exposure (given W) at a fixed level of the intermediate variable (Type 1 direct effect): $E(Y_{az} - Y_{0z}|W)$. In our example,

   $$\begin{aligned}\hat{E}(Y_{az} - Y_{0z}|W) &= \hat{E}(Y|A=1,W,Z=z) - \hat{E}(Y|A=0,W,Z=z) \quad (7)\\ &= 50 - 20W + 10z + 10zW\end{aligned}$$

   Equation (7) suggests that the direct effect of PI-based therapy depends on both the patient's viral load and treatment history.

3. Estimate an individual's Type 2 direct effect by replacing z in 7 with an estimate of the level of the intermediate 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.

   $$\hat{E}(Y_{1Z_0} - Y_{0Z_0}) = 50 - 20\hat{E}(W) + 10\hat{E}(Z_0) + 10\hat{E}(Z_0 * W) \tag{8}$$

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 antiretroviral regimen.

$$\hat{E}(Z|A,W) = 1.7 + 1.25W + 0.2A + 0.2AW \tag{9}$$

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

$$\hat{E}(W) = Pr(W=1) = 0.33 \tag{10}$$

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

$$\begin{aligned} \hat{E}(Z_0) &= \hat{E}(\hat{E}(Z|A=0,W)) \\ = \hat{E}(Z|A=0,W=1)\hat{Pr}(W=1) &+ \hat{E}(Z|A=0,W=0)\hat{Pr}(W=0) \\ &= 2.95 * 0.33 + 1.7 * 0.67 \\ &= 2.1 \end{aligned}$$

5. Similarly, the average of $E(Z|A=0,W)*W$ across the population provides an estimate of $E(Z_0 * W)$. In our example,

$$\begin{aligned} \hat{E}(Z_0 * W) &= \hat{E}(\hat{E}(Z|A=0,W)*W) \\ = \hat{E}(Z|A=0,W=1)(1)\hat{Pr}(W=1) &+ \hat{E}(Z|A=0,W=0)(0)\hat{Pr}(W=0) \\ &= 2.95 * 0.33 \\ &= 0.97 \end{aligned}$$

6. Substitute these values into model (8) to get an estimate of the Type 2 direct effect in the population. In our example, the direct effect of PI-based therapy on CD4 T cell count is estimated to be:

$$\begin{aligned} \hat{DE} &= \hat{E}(Y_{1Z_0} - Y_{0Z_0}) \\ &= 50 - 20\hat{E}(W) + 10\hat{E}(Z_0) + 10\hat{E}(Z_0 * W) \\ &= 50 - 20 * 0.33 + 10 * 2.1 + 10 * 0.97 \\ &= 74.1 \end{aligned}$$

We estimate that treatment of the study population with PI-based therapy vs. non-PI-based therapy would result in a 74 cell increase in average CD4 T-cells if the effect of PI-based therapy on viral load were blocked.

7 Assumptions.

Estimation of direct effects in the manner presented requires several assumptions. First, 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 (U_1 and U_2 in Figure 4), we assume that there are no unmeasured confounders of the effect of the exposure on the intermediate variable Z (U_3 in Figure 8). This assumption is necessary to ensure that regressing Z on the exposure and covariates (step 4) 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, as is illustrated in Figure 9, poor ability to adhere to prescribed medications results in both a higher viral load and an increased probability of assignment to a PI-based regimen, failure to include this variable 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.

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

$$A \perp Y_{az} | W \tag{11}$$

$$Z \perp Y_{az} | A, W \tag{12}$$

$$A \perp Z_a | W \tag{13}$$

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

$$E(Y_{az} - Y_{0z} | Z_0 = z, W) = E(Y_{az} - Y_{0z} | W) \tag{14}$$

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 fixed viral load.

Previous work discussing estimation of direct effects suggested that the assumptions necessary to make these effects identifiable were more restrictive than those presented here,[3,4,7] perhaps explaining in part why methodology for the estimation of direct effects has not been pursued further in the epidemiological literature. Robins and Greenland proposed an alternative to our direct effect assumption (14),[3] which states that the intermediate and the exposure of interest do not interact to affect outcome at the individual level.

$$Y_{az} - Y_{0z} \text{ is a random function } B(a) \text{ that does not depend on } z. \tag{15}$$

Such an assumption is both very restrictive and unrealistic in many biological settings. The assumption can be tested by examining the data; in our example, the presence of interactions between viral load (Z) and antiretroviral therapy (A) in model (6) suggests that the assumption is violated. Note that if the 'No Interaction' assumption were to hold, estimation of both Type 1 and Type 2 direct effects would simply require taking the average of model (7) (which would now not include z) across the population.

Pearl proposed a third alternative identifying assumption which states that,[7] 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:

$$Y_{a,z} \perp Z_0 | W \tag{16}$$

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 fixed 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 14 holds in essentially all cases where 18 holds, and in many cases where it does not.

We refer interested 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 (14) fails to hold, the method we present still estimates an interesting causal parameter: a summary of the direct effect of the exposure in the population, with the intermediate fixed at its mean counterfactual level in the absence of exposure. As a result, we feel that the identifiability assumption should not present a barrier to researchers interested in the estimation of direct effects.

8 Discussion.

The estimation of direct effects is a common goal in epidemiologic research. In the estimation of direct effects, as 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 fixed level defined by the researcher (Type 1 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 were blocked, allowing the intermediate to follow the course it would have taken in the absence of exposure (Type 2 direct effect), multivariable regression alone may be insufficient. In the case where exposure and intermediate do not interact at the individual level to affect outcome, Type 1 and Type 2 direct effects are equivalent; however, such an assumption is not required to ensure their identifiability.

We have presented a straightforward method for estimating Type 2 direct effects and illustrated it in a simple single time-point setting where exposure and intermediate interact to cause disease. Our method involves fitting a multivariable regression of outcome on exposure, confounders, and intermediate, and an additional multivariable regression of the intermediate on confounders and exposure. In settings where a confounder is affected by the exposure, as well as in longitudinal settings, the same general approach can be used, but methods other than multivariable regression (such as inverse probability weighting or g-computation) must be used to estimate $E(Y_{az} - Y_{0z}|W)$. We hope that researchers will be encouraged to estimate direct effects of interest and, where appropriate, to continue their analyses beyond fitting multivariable regression models of the outcome.

9 Appendix.

9.1 Comparison with identifying assumptions in prior literature

We propose the following novel assumption for identification of Type 2 direct causal effects:

$$E[Y_{az} - Y_{0z}|Z_0, W] = E[Y_{az} - Y_{0z}|W] \text{ for all } a \text{ and } z, \tag{17}$$

Comparison with Pearl: Pearl shows (using the structural equation framework) that the Type 2 direct effect is identifiable,[7] if

$$Y_{az} \perp Z_0 \mid W \text{ for all } z, \tag{18}$$

It is of interest to compare our assumption (17) with (18).

An alternative way of formulating this assumption (18) is that there exists a function m such that

$$Y_{a,z} = m(a, z, W, e), \tag{19}$$

where e is a random variable which is conditionally independent of Z_0, given W. Stated in words, Pearl assumes that,[7] 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 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 (18) holds at two 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 (17) 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 (18) holds, one will also have that our assumption holds. On the other hand, it is easy to construct examples in which our assumption holds, while (18) fails to hold.[13] We refer to Robins for further discussion of the limitations of assumption (18).[4]

Comparison with Robins:

Robins and Greenland propose an alternative identifying assumption,[3] which they call the No-Interaction Assumption:

$$Y_{az} - Y_{0z} \text{ is a random function } B(a) \text{ that does not depend on } z. \qquad (20)$$

In words, this assumption states that the individual 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.[3] The "No-interaction Assumption" implies, in particular, that $EY_{az} = m_1(a) + m_2(z)$ 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 as well.

Applied to our HIV example, Robins' assumption 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 were controlled at a high level (the study population was virologically failing) or controlled at a low level (the study population was virologically succeeding). 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. Thus, we would expect that, if PI-based therapy has an anti-apoptotic direct effect on CD4 T-cell count (i.e., 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.

9.2 Identifiability result

Under the direct effect assumption (17), we have the following identifiability result, presented as a theorem:

Theorem 1 *Let $DE(a) = E(Y_{aZ_0} - Y_{0Z_0})$. Assume that (17) holds. Then,*

$$\begin{aligned} DE(a) &= \widetilde{DE}(a) \\ &\equiv E_W \int \{E(Y_{az} \mid W) - E(Y_{0z} \mid W)\} \, dF_{Z_0|W}(z). \end{aligned} \tag{21}$$

We provide here the formal identifiability result for our theorem (21) based on our assumption (17),

$$\begin{aligned} \text{DE} &= E(Y_{aZ_0} - Y_{0Z_0}) \\ \text{DE} &= E_W(E(Y_{aZ_0} - Y_{0Z_0} \mid W)) \\ \text{DE} &= E_W(E_{Z_0|W}(E((Y_{aZ_0} - Y_{0Z_0} \mid Z_0, W))) \\ \text{DE} &= E_W \int_{\mathcal{Z}} E(Y_{az} - Y_{0z} \mid Z_0 = z, W) dF_{Z_0|W}(z) \\ \text{DE} &= E_W \int_{\mathcal{Z}} E(Y_{az} - Y_{0z}|W) dF_{Z_0|W}(z) \text{ by (17)} \\ \text{DE} &\equiv \widetilde{\text{DE}}, \end{aligned}$$

where the right-hand 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 (18) also implies $\text{DE} = \widetilde{\text{DE}}$. Thus Pearl's identifiability mapping is the same as ours (21), but it was based on a different assumption. Similarly, under the assumption of Robins (20) we have $Y_{aZ_0} - Y_{0Z_0} = Y_{az} - Y_{0z}$ for any z so that

$$E(Y_{aZ_0} - Y_{0Z_0}) = E(Y_{az} - Y_{0z}), \tag{22}$$

where the latter quantity does not depend on z. Robins' identifiability mapping (22) corresponds with ours using an empty W (and thus with Pearl's). since the integration w.r.t. F_{Z_0} does not affect the integral. We conclude that all three identifiability mappings agree with each other (except that Robins avoids integration w.r.t. $F_{\bar{Z}_0}$ by making the "No-Interaction assumption"), but that the model assumptions which were used to validate the identifiability mapping are different. Our result shows that the identifiability mapping of Pearl holds under a much less restrictive *union-assumption*: that is, the identifiability result presented in Theorem 1 holds if either our assumption holds, or the (18) assumption holds, or the "No-Interaction Assumption" holds.

9.3 Estimation approach in single time point case

Consider the simple single time-point data structure W, A, Z, Y. For simplicity, we assume that all variables are univariate. We need to 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 assumption (17)). 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 direct effect.

Under the assumption that (A, Z) is randomized w.r.t. 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

$$E(Y \mid A, Z, W) = A(\beta_0 + \beta_1 Z + \beta_2 W + \beta_3 ZW) + (\alpha_0 + \alpha_1 Z + \alpha_2 W + \alpha_3 ZW),$$

so that we have the model

$$E(Y_{az} - Y_{0z} \mid W) = a(\beta_0 + \beta_1 z + \beta_2 W + \beta_3 zW). \tag{23}$$

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

In order to estimate the direct effect, we need to take the conditional expectation of (23) over z w.r.t. 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 (23) in z, it follows that the direct effect is now modeled as

$$DE = A(\beta_0 + \beta_1 E(m(0, W \mid \lambda)) + \beta_2 EW + \beta_3 E(m(0, W \mid \lambda)W)),$$

where $m(0, W|\lambda)$ 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.

10 Acknowledgements

We thank James Robins and Rudolphe Thiebaut for helpful comments. Maya Petersen is supported by a Pre-Doctoral Fellowship from the Howard Hughes Medical Institute.

11 References

1. Valimaki M, Tiitinen A, Ylikorkala O, Evio S. Effects of hormone therapy and alendronate on C-reactive protein, e-selectin, and sex hormone-binding globulin in osteoporotic women. *Fertil Steril.* 2003: 80(3):541-545.

2. Pearl J, Greenland S, Robins JM. Causal diagrams for epidemiologic research. *Epidemiology.* 1999; 10(1):37-81.

3. Robins JM, Greenland S. Identifiability and exchangeability for direct and indirect effects. *Epidemiology.* 1992; 3(0):143-155.

4. Robins JM. Semantics of causal DAG models and the identification of direct and indirect effects. In: N. Hjort P. Green and S. Richardson, 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 Epidemiology.* 2000; 151:S52.

7. Pearl J. *Causality: Models, Reasoning, and Inference.* Cambridge: Cambridge University Press; 2000.

8. Robins JM, Hernan MA, and Brumback B. Marginal structural models and causal inference in epidemiology. *Epidemiology.* 2000; 11(5):550-560.

9. Martinez-Picado J, Savara AV, Sutton L, D'Aquila R. Replicative fitness of protease inhibitor-resistant mutants of Human Immunodeficiency Virus type-1. *J Virol.* 1999; 73(5):3744-3752.

10. Deeks SG, Wrin T, Liegler T, et al. Virologic and immunologic consequences of discontinuing combination antiretroviral-drug therapy in HIV-infected patients with detectable viremia. *N. Engl. J. Med..* 2000; 344(7):472-480.

11. Deeks SG, Barbour J, Martin J, Swanson M, Grant R. Sustained CD4+ T cell response after virologic failure of protease inhibitor-based regimens in patients with Human Immunodeficiency Virus infection. *J Infect Dis.* 2000; 181(3):946-53.

12. Phenix B, Angel J, Mandy F, Kravcik S. Decreased HIV-associated T-cell apoptosis by HIV protease inhibitors. *AIDS Res Hum Retroviruses.* 2000; 16(6):559-67.

13. van der Laan MJ and 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

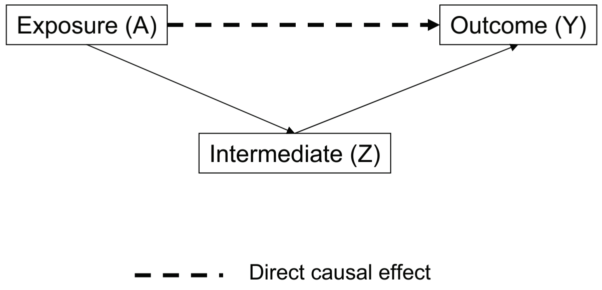

Figure 1: Basic causal structural of direct effect questions
Example 1: A=type of antiretroviral therapy (PI-based or not), Z=viral load, Y=CD4 T-cell count
Example 2: A=pollution level, Z=use of rescue medication, Y=lung function
Example 3: A=hormone therapy, Z=C-reactive protein, Y=cardiovascular event

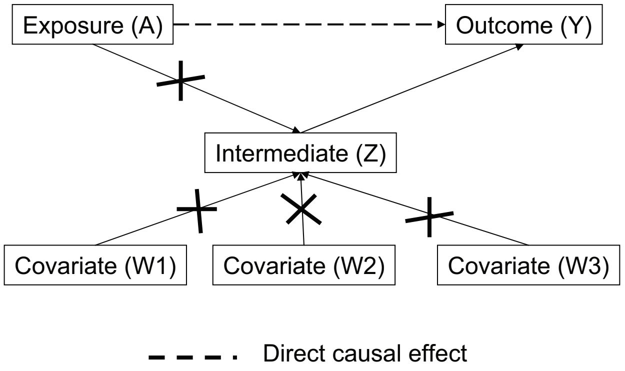

Figure 2: Type 1 Direct Effect of A on Y, holding Z at a fixed level (blocking all effects on Z).

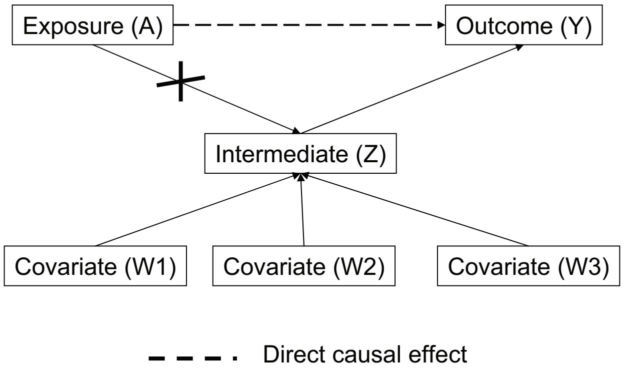

Figure 3: Type 2 Direct Effect of A on Y, blocking only the effect of A on Z.

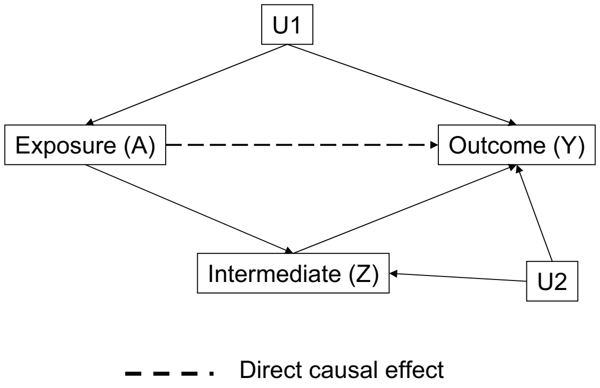

Figure 4: Unmeasured confounders of exposure effect (U1) and intermediate effect (U2).

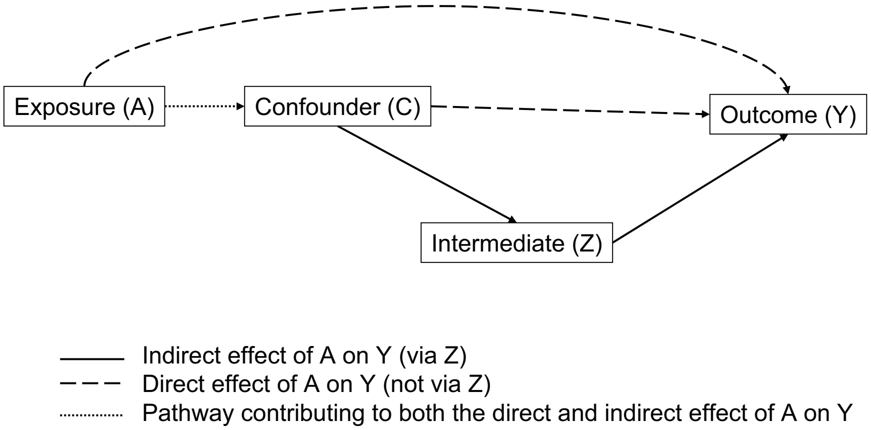
Exposure (A)
Confounder (C)
Outcome (Y)
Intermediate (Z)
Indirect effect of A on Y (via Z)
Direct effect of A on Y (not via Z)
Pathway contributing to both the direct and indirect effect of A on Y

Figure 5: Confounding by a causal intermediate.

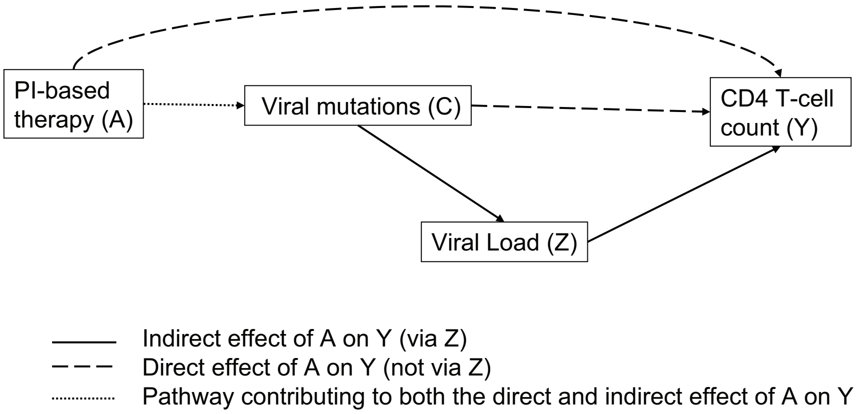
PI-based therapy (A)
Viral mutations (C)
CD4 T-cell count (Y)
Viral Load (Z)
Indirect effect of A on Y (via Z)
Direct effect of A on Y (not via Z)
Pathway contributing to both the direct and indirect effect of A on Y

Figure 6: Confounding by a causal intermediate: Illustration based on Example 1.

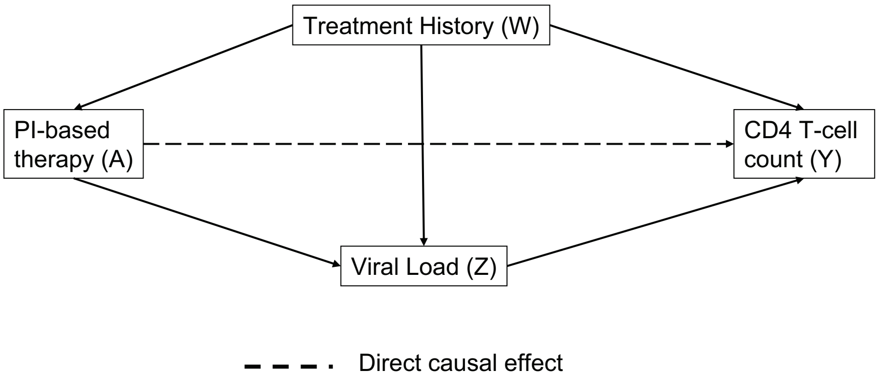

Figure 7: Causal structure for numerical example, based on Example 1.

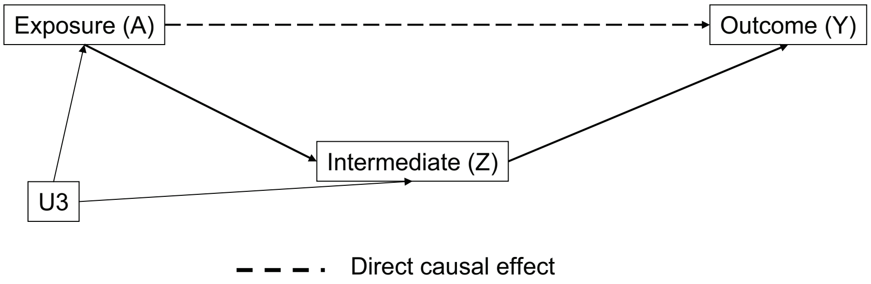

Figure 8: Confounding of the effect of exposure on the intermediate variable.

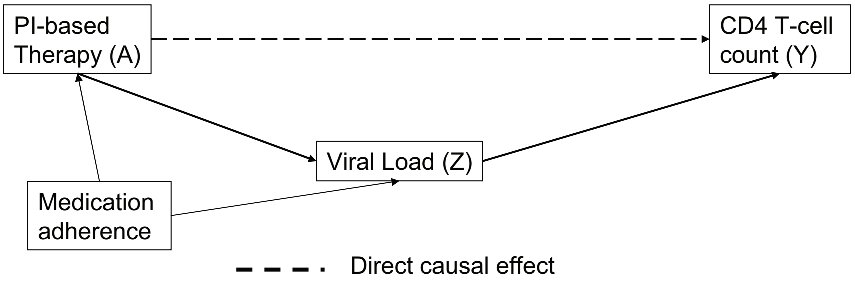

Figure 9: Confounding of the effect of exposure on the intermediate variable: Illustration based on Example 1.