University of California, Berkeley
U.C. Berkeley Division of Biostatistics Working Paper Series

Year 2004 *Paper* 155

Estimation of Direct and Indirect Causal Effects in Longitudinal Studies

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

*Division of Biostatistics, School of Public Health, University of California, Berkeley, laan@berkeley.edu

†Division of Biostatistics, School of Public Health, University of California, Berkeley, mayaliv@hotmail.com

http://biostats.bepress.com/ucbbiostat/paper155

Estimation of Direct and Indirect Causal Effects in Longitudinal Studies

Mark J. van der Laan and Maya L. Petersen

Abstract

The causal effect of a treatment on an outcome is generally mediated by several intermediate variables. Estimation of the component of the causal effect of a treatment that is mediated by a given intermediate variable (the indirect effect of the treatment), and the component that is not mediated by that intermediate variable (the direct effect of the treatment) is often relevant to mechanistic understanding and to the design of clinical and public health interventions. Under the assumption of no-unmeasured confounders, Robins & Greenland (1992) and Pearl (2000), develop two identifiability results for direct and indirect causal effects. They define an individual direct effect as the counterfactual effect of a treatment on an outcome when the intermediate variable is set at the value it would have had if the individual had not been treated, and the population direct effect as the mean of these individual counterfactual direct effects. The identifiability result developed by Robins & Greenland (1992) relies on an additional "No-Interaction Assumption", while the identifiability result developed by Pearl (2000) relies on a particular assumption about conditional independence in the population being sampled. Both assumptions are considered very restrictive. As a result, estimation of direct and indirect effects has been considered infeasible in many settings. We show that the identifiability result of Pearl (2000), also holds under a new conditional independence assumption which states that, within strata of baseline covariates, the individual direct effect at a fixed level of the intermediate variable is independent of the no-treatment counterfactual intermediate variable. We argue that our assumption is typically less restrictive than both the assumption of Pearl (2000), and the "No-interaction Assumption" of Robins & Greenland (1992). We also generalize the current definition of the direct (and indirect) effect of a treatment as the population mean of individual counterfactual direct (and indirect) effects to 1) a general parameter of the population distribution of individual counterfactual

direct (and indirect) effects, and 2) change of a general parameter of the population distribution of the appropriate counterfactual treatment-specific outcome. Subsequently, we generalize our identifiability result for the mean to identifiability results for these generally defined direct effects. We also discuss methods for modelling, testing, and estimation, and we illustrate our results throughout using an example drawn from the treatment of HIV infection.

1 Introduction.

Consider a longitudinal study in which one collects on each randomly sampled subject the chronological data structure $Z(0), L(0), A(0), \ldots, Z(K), L(K), A(K), Z(K+1), Y = L(K+1)$, where $L(j)$ is a time-dependent covariate measured at time j, $A(j)$ is a time-dependent treatment, $Z(j)$ is a time-dependent covariate of interest, $j = 0, \ldots, K+1$, and Y is the final outcome of interest measured at time $K+1$, which denotes the end of the study. Let $W \equiv (Z(0), L(0))$ denote the baseline covariates measured before the assignment of the initial treatment $A(0)$. The simplest special case (corresponding with $K = 0$) of this data structure is $O = (W, A, Z, Y)$, where W denotes baseline covariates, A denotes treatment, Z denotes an intermediate outcome, and Y is the final outcome of interest. In this article we are concerned with answering questions such as "What is the direct causal effect of A on Y?" and "What is the indirect causal effect of A on Y through Z?".

We will rely throughout the paper on an example drawn from our research on the treatment of Human Immunodeficiency Virus (HIV) to illustrate our notation and results. Antiretroviral therapy suppresses the replication of HIV, reflected in a reduced plasma HIV RNA level (viral load). As a result of reduced viral replication, a patient's CD4 T-cell count increases, restoring immunologic function. Antiretroviral therapy may also increase CD4 T-cell counts in ways not mediated by changes in viral load(Deeks et al. (2000)). This example is illustrated in Figure 1. We are interested in the question "What is the direct causal effect of antiretroviral therapy on CD4 T- cell count (not mediated by changes in viral load)?". Identifying such a direct effect would have important implications for understanding both the mechanics of antiretroviral action and the appropriate clinical response to viral resistance, which can reduce or eliminate the effect of treatment on viral load.

Using the notation described above, in our example A denotes antiretroviral therapy, Z denotes viral load, Y denotes CD4 T-cell count at the end of the study, and W denotes any baseline covariates (such as age, sex, injection drug use status, etc...). For simplicity, for the majority of the paper we treat antiretroviral therapy as binary (treated or not) and assume that therapy, viral load, outcome and covariates are each measured at a single point in time. However, the example can be easily generalized to more complex data structures.

Under the assumption of no-unmeasured confounders, Robins and Green-

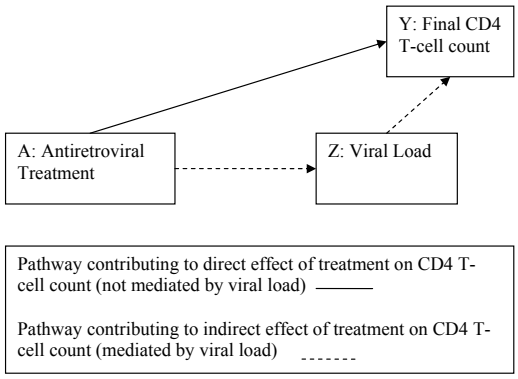

Figure 1:

land (1992) and Pearl (2000) develop two identifiability results for direct (and indirect) causal effects. They define an individual direct effect as the counterfactual effect of a treatment on an outcome when the intermediate variable is set at the value it would have had if the individual had not been treated. The population direct effect is defined as the mean of these individual counterfactual direct effects. The identifiability result developed by Robins and Greenland (1992) relies on an additional "No-Interaction Assumption", while the identifiability result developed by Pearl (2000) relies on a particular assumption about conditional independence in the population being sampled. Both assumptions are considered very restrictive and unrealistic. As a result, estimation of the direct and indirect effects of treatment has been considered infeasible in many settings. We show that the identifiability result of Pearl (2000) also holds under a new conditional independence assumption which states that, within strata of baseline covariates, the individual direct causal effect at a fixed level of the intermediate variable is independent of the no-treatment counterfactual intermediate variable. Using both theoretical arguments and the HIV example presented, we discuss the interpretation and plausibility of our assumption in comparison to the assumptions of Robins and Greenland (1992) and Pearl (2000). We also discuss methods for modelling, testing, and estimation.

We also generalize the current definition of the direct (and indirect) effect of a treatment as the population mean of individual counterfactual direct (and indirect) effects to 1) a general parameter of the population distribution of individual counterfactual direct (and indirect) effects, and 2) change of a general parameter of the population distribution of the appropriate counterfactual treatment-specific outcome. Subsequently, we generalize our identifiability result for the mean to identifiability results for these generally defined direct effects. One could name these two categories of direct effects as "parameter of change" and "change of parameter", respectively, where we note that in the special case of the mean these two definitions agree with each other.

This article is organized as follows. In Section 2 we review the statistical counterfactual framework as used by Robins and Greenland (1992), Robins (2003), and present their definitions of direct and indirect effects on the mean. In Section 3 we present our new conditional independence assumption and the corresponding identifiability results for direct causal effects. We also provide a detailed discussion of our assumption and comparison with the previously used assumptions for identifying the direct and indirect effect on the mean.

In Section 4 we show how one can estimate these direct causal effects with current statistical methods based on marginal structural models, as detailed in Robins (1997) and van der Laan and Robins (2002). In this section we also present a modelling strategy and corresponding test of the null hypothesis of "no direct effect". In Section 5 we generalize the definition of direct and indirect effects, and generalize our identifiability results. Finally, in Section 6 we discuss the extensions of our identifiability result to general longitudinal data structures, including the case that the outcome of interest is a survival time.

2 Definition of a direct effect: Review

We follow the statistical framework and definitions of direct and indirect effects as presented in Robins and Greenland (1992) and Robins (2003). This statistical framework represents the observed data as a missing data structure, where the full data is a collection of counterfactual data structures corresponding with set values of the treatment and intermediate variables. In other words, in the full data we would observe, for each individual, the value of the intermediate variable over time resulting from each possible treatment history, and the value of the covariate process over time, including the outcome, resulting from each combination of possible treatment history and possible intermediate variable history. Instead, our observed data is only a subset of this full data, consisting of a single treatment history and the corresponding intermediate variable and covariate processes.

2.1 Causal effects of joint treatment and intermediate covariate process.

Given a time-dependent process $X(t)$, we will adopt the notation $\bar{X}(t) = (X(s) : s \leq t)$.

Counterfactuals only controlling treatment: Let $\bar{A} = \bar{A}(K) = (A(0), \ldots, A(K))$ denote the multivariate treatment regime, or the treatment history through time K, and let $\mathcal{A}$ denote all possible treatment histories. For each $\bar{a} \in \mathcal{A}$, let $X_{\bar{a}}(j) \equiv (Z_{\bar{a}}(j), L_{\bar{a}}(j))$ denote the treatment-specific process one would have observed if the subject would have followed treatment regime $\bar{A} = \bar{a}$, $j = 0, \ldots, K+1$. It is assumed that $X_{\bar{a}}(j) = X_{\bar{a}(j-1)}(j)$; in other words, the values of the intermediate variable and covariates are not affected by treat-

ment that occurs after they are measured (we assume that $A(j)$ is measured after $X(j)$). It is further assumed that the observed $X(j) = (Z(j), L(j))$ equals the treatment specific $(Z_{\bar{a}}(j), L_{\bar{a}}(j))$ corresponding with the treatment the subject actually took: that is, $L(j) = L_{\bar{A}(j-1)}(j)$, and $Z(j) = Z_{\bar{A}(j-1)}(j)$, $j = 0, \ldots, K+1$. Under this so called consistency assumption, we have the following relation between the collection of counterfactual data structures and the observed data:

$$O = (\bar{A}(K), \bar{L}_{\bar{A}}(K+1)).$$

In the special case $O = (W, A, Z, Y)$ we have $O = (W, A, Z_A, Y_A)$, where (Z_a, Y_a) represents the treatment-specific counterfactual outcome of (Z, Y).

In our example, $\mathcal{A}$ denotes all possible antiretroviral treatments, and consists of $a = 0$ (untreated) and $a = 1$ (treated). The counterfactual viral load and CD4 T-cell count that would have been observed under no treatment is denoted $X_0 \equiv (Z_0, Y_0)$, and the viral load and CD4 T-cell count that would have been observed under treatment is denoted $X_1 \equiv (Z_1, Y_1)$. We assume that the viral load and CD4 count we observe for a treated subject are equivalent to the subject's counterfactual viral load and CD4 T-cell count under treatment, and that the viral load and CD4 T-cell count we observe for an untreated subject are equivalent to the counterfactual viral load and CD4 T-cell count under no treatment (the consistency assumption). Observed data for a given subject can then be represented as the subject's observed treatment, the counterfactual CD4 T-cell count and viral load under the observed treatment, and any additional baseline covariates W.

Definition of a causal effect of treatment: One can now define a causal effect of treatment on (e.g.) the outcome Y in terms of a particular difference between the distribution of $Y_{\bar{a}}$ and the distribution of Y_0. Applying this definition to our example, the causal effect of antiretroviral treatment is defined as the difference in the distribution of final CD4 T-cell count that would have been observed if our study population were treated and the distribution of CD4 T-cell count that would have been observed if our study population were untreated. The causal effect of a treatment can be further defined as conditional on baseline covariates $V \subset L(0)$ (for example, marginal structural models (Robins (1997), Robins (2000), van der Laan and Robins (2002) define a causal effect in this manner), or on an observed past (for example, structural nested mean models (Robins (1989), Robins (1994) define a casual effect in this manner).

Alternatively, one can define the causal effect of treatment as a particular parameter of the distribution of individual causal effects $Y_{\bar{a}} - Y_0$. For example, structural nested models focus on modelling the distribution of individual causal effects (Robins (1997), van der Laan and Robins (2002)). Applied to our example, the causal effect of antiretroviral treatment is defined as some parameter (such as the mean or median) of the individual difference in CD4 T-cell count that would have been observed for each subject in the study if he/she had been treated vs. untreated.

Identifiability of the counterfactual treatment-specific distributions: The distribution of $\bar{X}_{\bar{a}} = (\bar{Z}_{\bar{a}}, \bar{Y}_{\bar{a}})$ is identified by the G-computation formula (or Inverse Probability of Treatment Weighted, or IPTW, and Double-Robust Inverse Probability of Treatment Weighted, or DR-IPTW, estimating function) under the assumption that $A(j)$ is sequentially randomized (SRA) and that experimental treatment assumption (ETA) holds for the conditional density of $A(j)$, given the observed past. The SRA assumption states here that $A(j)$ is conditionally independent of $X = (\bar{X}_{\bar{a}} : \bar{a} \in \mathcal{A})$, given the observed past $(\bar{A}(j-1), \bar{X}_{\bar{A}(j-1)}(j))$, $j = 0, \ldots, K+1$. In our example, the SRA states that whether an individual is treated or not is independent of his/her counterfactual CD4 T-cell count and viral load, given observed covariates, or in other words, there are no unmeasured variables that predict both treatment and CD4 T-cell count, or both treatment and viral load (i.e., no unmeasured confounders). We refer to van der Laan and Robins (2002), Gill and Robins (2001), Yu and van der Laan (2002) for a definition of the experimental treatment assignment assumption, and the precise statement of the identifiability result based on the G-computation formula for discrete as well as continuous data.

Counterfactuals controlling both treatment and intermediate covariate process: The definition of direct and indirect causal effects requires, beyond the definition of treatment-specific counterfactual processes $\bar{Z}_{\bar{a}}$, the definition of counterfactuals for Y in which $\bar{Z}$ is also controlled. That is, one views $(\bar{A}, \bar{Z})$ as a joint treatment process which can potentially be controlled by the experimenter. Let $\mathcal{Z}$ denote the set of all possible values for $\bar{Z}$. For each $\bar{a} \in \mathcal{A}$ and $\bar{z} \in \mathcal{Z}$, let $\bar{L}_{\bar{a},\bar{z}} \equiv (L_{\bar{a},\bar{z}}(j) : j = 0, \ldots, K+1)$ denote the treatment-specific L-process one would have observed if the subject would have followed treatment regime $\bar{A} = \bar{a}$, and if his/her Z-process would have been controlled at $\bar{Z} = \bar{z}$.

It is assumed that $L_{\bar{a},\bar{z}}(j) = L_{\bar{a}(j-1),\bar{z}(j-1)}(j)$. In other words, the covariate process is not affected by either treatment or level of the intermediate variable

occurring after the covariate process is measured. In addition, it is assumed that the observed $L(j)$ equals the treatment-specific $L_{\bar{a},\bar{z}}(j)$ corresponding with the treatment and covariate process the subject actually followed: that is, $L(j) = L_{\bar{A}(j-1),\bar{Z}(j-1)}(j)$, $j = 0, \ldots, K+1$. Under this latter consistency assumption, and the previous definition of $\bar{Z}_{\bar{a}}$ and corresponding consistency assumption $\bar{Z} = \bar{Z}_{\bar{A}}$, we have the following relation between the complete collection of counterfactual data structures $X \equiv (\bar{Z}_{\bar{a}}, \bar{L}_{\bar{a},\bar{z}} : \bar{a}, \bar{z})$ (i.e., the full data structure one would have liked to observe) and the observed data structure:

$$O = (\bar{A}(K), \bar{Z} = \bar{Z}_{\bar{A}}(K+1), \bar{L}_{\bar{A},\bar{Z}}(K+1)).$$

In particular, $Y_{\bar{a},\bar{z}} = L_{\bar{a},\bar{z}}(K+1)$ denotes the counterfactual outcome one would have observed if the subject would have followed treatment regime $\bar{A} = \bar{a}$, and if his/her Z-process would have been controlled at $\bar{Z} = \bar{z}$. The counterfactuals $Y_{\bar{a}}$ and $Y_{\bar{a},\bar{z}}$ can be related to each other by assuming $Y_{\bar{a}} = Y_{\bar{a},\bar{Z}_{\bar{a}}}$, and, more general, $\bar{L}_{\bar{a}} = \bar{L}_{\bar{a},\bar{Z}_{\bar{a}}}$.

Applying this notation to our example, we define $Y_{a,z}$ as the counterfactual outcome (CD4 T-cell count) that would have been observed if the individual had followed antiretroviral treatment regime a, and had his/her viral load controlled at z. The full data for a given individual are composed of baseline covariates and the counterfactual outcomes (CD4 T-cell counts) under possible viral loads and treatments and the viral loads under possible treatments. In the observed data, only one viral load can be observed for any given individual. Further, under the consistency assumption, the viral load and CD4 T-cell count observed are the counterfactual viral load and CD4 T-cell count under the observed treatment; both viral load and CD4 T-cell count observed for a given individual will correspond to

the counterfactual viral load and CD4 T-cell count for the treatment the subject actually took.

Identifiability of counterfactual joint "treatment" -specific distributions: The distribution of $L_{\bar{a},\bar{z}}$ is identified (e.g., by the G-computation formula) from the data if the joint "treatment" $(A(j), Z(j))$ is conditionally independent of the collection of counterfactuals X, given the observed past $\bar{A}(j-1), \bar{Z}(j-1), \bar{L}(j-1)$, $j = 1, \ldots, K$, and an experimental joint treatment assignment assumption (ETA) holds for the conditional density of $(A(j), Z(j))$, given the observed past (see van der Laan and Robins (2002)). The first assumption is referred to as the sequential randomization assumption (SRA) of the treatment $(A(j), Z(j))$. If one is only concerned with

identifiability of the distribution of $Y_{\bar{a},\bar{z}}$, then this sequential randomization assumption can be weakened to just assuming that $(A(j), Z(j))$ is conditionally independent of the collection of counterfactuals $(Y_{\bar{a},\bar{z}} : \bar{a}, \bar{z})$, given the observed past $\bar{A}(j-1), \bar{Z}(j-1), \bar{L}(j-1)$, $j = 1, \ldots, K$. In our example, this is equivalent to assuming that, within subgroups defined by baseline covariates W, there are no unmeasured variables that predict both antiretroviral treatment and viral load, as well as the outcome, CD4 T-cell count.

2.2 Direct and indirect causal effects.

Robins and Greenland (1992)), Robins (2003), and Pearl (2000) provide the following definition of a direct effect

$$\mathrm{DE}(\bar{a}) = E(Y_{\bar{a},\bar{Z}_0} - Y_{0,\bar{Z}_0}). \tag{1}$$

We refer to $Y_{\bar{a},\bar{Z}_0} - Y_{0,\bar{Z}_0}$ as the individual direct effect. A direct effect is thus defined as the population mean of individual direct effects, where an individual direct effect is the difference between the outcome when an individual is treated and the intermediate variable is set at its value under no treatment, and the outcome when the same individual is not treated. In our example, we could calculate an individual's direct effect using the full data by taking the difference between the individual's counterfactual CD4 T-cell count under treatment, with viral load set to its untreated value, and the same individual's CD4 T-cell count under no treatment. The direct effect for the study population would then be calculated by taking the mean of these individual direct effects.

Similarly, Robins and Greenland (1992)), Robins (2003), and Pearl (2000) define an indirect effect as

$$\mathrm{IDE}(\bar{a}) = E\left(Y_{\bar{a},\bar{Z}_{\bar{a}}} - Y_{\bar{a},\bar{Z}_0}\right). \tag{2}$$

We also note that (see e.g., Robins (2003))

$$\begin{aligned} TE &\equiv E(Y_{\bar{a}} - Y_0) = E(Y_{\bar{a},\bar{Z}_{\bar{a}}} - Y_{0,\bar{Z}_0}) \\ &= E(Y_{\bar{a},\bar{Z}_{\bar{a}}} - Y_{\bar{a},\bar{Z}_0}) + E(Y_{\bar{a},\bar{Z}_0} - Y_{0,\bar{Z}_0}) \\ &= IDE + DE. \end{aligned} \tag{3}$$

In order to avoid too much repetition, in this article we will focus on testing and estimation of direct causal effects. A similar approach can be followed for inference regarding indirect causal effects.

Note the difference between $E(Y_{\bar{a},\bar{Z}_0} - Y_{0,\bar{Z}_0})$ and $E(Y_{\bar{a},\bar{z}} - Y_{0,\bar{z}})$. The former expression evaluates the mean of the treatment effect in the population when the intermediate variable follows the trajectory it would have had for each individual under no treatment. The latter expression evaluates the mean of the treatment effect in the population when the intermediate variable is set to a constant pre-specified level/trajectory for all the individuals in the population.

In some settings, particularly if the intermediate variable $\bar{z}$ is amenable to public health intervention, $E(Y_{\bar{a},\bar{z}} - Y_{0,\bar{z}})$ may be a parameter of interest in its own right. For example, a physical exercise program may reduce heart disease both directly and because it encourages individuals to stop smoking. One might be interested in asking: "What would be the effect of exercise on heart disease if the whole population were to stop smoking?". Such a question could be addressed by evaluating $E(Y_{\bar{a},\bar{z}} - Y_{0,\bar{z}})$, treating exercise and smoking as a joint treatment, and would not require any additional assumptions beyond the SRA and ETA to be identifiable from the observed data.

In other settings, however, $E(Y_{\bar{a},\bar{z}} - Y_{0,\bar{z}})$ is less interesting. In our HIV example, the question "What would the effect of antiretroviral treatment be if viral load in the population were controlled at a specified level?" has no meaningful clinical or public health interpretation; no interventions other than antiretroviral treatment are available to control viral load. However, it is still interesting to ask the hypothetical mechanistic question "What would the effect of antiretroviral treatment on CD4 T-cell count be if treatment had no effect on viral load; in other words, if viral load remained at the level it would have had for each individual in the absence of treatment?". The definition of direct effect proposed by Robins and Greenland (1992)), Robins (2003), and Pearl (2000), which treats the intermediate variable as an additional counterfactual random variable, addresses this type of question.

A direct (and indirect) causal effect, as defined here, cannot be identified from the observed data without making additional non-testable assumptions beyond the SRA and ETA. The main cause of this lack of identifiability is the fact that, given the full collection of counterfactuals $X = (Y_{\bar{a},\bar{z}}, Z_{\bar{a}} : \bar{a}, \bar{z})$ for a given subject, evaluation of $Y_{\bar{a},Z_0}$ requires combining counterfactuals under two different treatment regimes: that is, one first selects the counterfactual Z_0 corresponding with *no-treatment*, and subsequently, one selects the counterfactual $Y_{\bar{a},\bar{z}}$ corresponding with $\bar{z} = Z_0$, but *with treatment* $\bar{a}$. Since one never observes these two counterfactuals simultaneously, any parameter of

this distribution is non-identifiable without making additional assumptions (Robins (2003)).

To illustrate the need for additional assumptions to ensure the identifiability of direct effects, as compared to total effects, consider our example. As stated above, our full data consist of the counterfactual outcomes (CD4 T-cell counts) under all possible treatments and viral loads, the counterfactual viral load under all possible treatments, and baseline covariates ($X \equiv (Y_{a,z}, Z_a, W : a, z)$). From this full data, we can compute the two counterfactual outcomes used to define the direct effect: CD4 T-cell count if the individual was untreated, and viral load was controlled at its untreated value ($Y_{0,Z_0} = Y_0$), and CD4 T-cell count if the individual was treated, and viral load was controlled at its untreated value (Y_{1,Z_0}). As in any causal inference problem, we only observe the counterfactual outcome for a given individual under a single treatment, Y_1 or Y_0. Estimation of the direct effect of treatment is further complicated by the fact that we *never* observe one of the counterfactual outcomes, (Y_{1,Z_0}) used to define the direct effect, because this counterfactual outcome corresponds to two different treatments in the same individual.

3 Identifiability result for direct causal effect.

In the next subsections we will present the corresponding formal identifiability result, provide understanding of our proposed conditional independence assumption (4), and compare it with the assumptions proposed in the current literature.

3.1 Identifying assumption for direct causal effect.

We propose the following assumption for identification of the direct causal effect:

$$Y_{\bar{a},\bar{z}} - Y_{0,\bar{z}} \perp \bar{Z}_0 \mid W \text{ for all } \bar{a} \text{ and } \bar{z}, \tag{4}$$

We will refer to $Y_{\bar{a},\bar{z}} - Y_{0,\bar{z}}$ as the individual direct effect at a fixed $\bar{z}$. In words, this assumption states that, within strata of baseline covariates, the direct effect of treatment at a fixed level of the intermediate variable does not depend on an individual's counterfactual level of the intermediate variable under no treatment.

An alternative way of formulating our assumption (4) is that there exists a function m such that

$$Y_{\bar{a},\bar{z}} = Y_{0,\bar{z}} + m(\bar{a}, \bar{z}, W, e), \tag{5}$$

where e is a random variable which is conditionally independent of $\bar{Z}_0$, given W. This assumption would hold, in particular, if e is an exogenous variable independent of the subject. We note that this assumption puts no constraints on $E(Y_{\bar{a},\bar{z}} \mid W)$, and it is a non-testable assumption. For example, one might have that $Y_{\bar{a},\bar{z}} = Y_{0,\bar{z}} + g(\bar{a}, \bar{z}, W) + e$, where the variations e are not predictive of the subject's progression of disease (as measured by $\bar{Z}_0$) under no-treatment.

In words, this assumption states that the individual direct causal effect $Y_{\bar{a},\bar{z}} - Y_{0,\bar{z}}$ is a deterministic function of $\bar{a}, \bar{z}$, the baseline covariates W, and an "exogenous" error. That is, within a subpopulation defined by a strata of W, the variation of $Y_{\bar{a},\bar{z}} - Y_{0,\bar{z}}$ among subjects is completely explained by random fluctations independent of the subject's characteristics related to $\bar{Z}_0$. In order to make this a reasonable assumption, it is important that one measures enough baseline covariates explaining this variation in direct effects at a fixed $\bar{z}$.

In the context of the HIV example, we assume that, within strata of baseline covariates, the direct effect of antiretroviral treatment on CD4 T-cell count, controlling viral load at a fixed level z, does not depend on what an individual's viral load would have been under no treatment. For example, if one were to control viral load at a high level among a group of individuals with identical baseline covariates, the effect of treatment on CD4 T-cell count would not vary based on an individual's viral load under no treatment. Thus, we must include in our baseline covariates any variables that are associated with viral load under no treatment and also predict the magnitude of the individual direct effect at a fixed viral load. Note that, in our example, $Y_{a,z} - Y_{0,z}$ is a hypothetical construct; as discussed above, there is no intervention that allows us to control viral load while changing antiretroviral treatment.

Under this assumption, we have

$$\begin{aligned} \text{DE} &= E_W E(Y_{\bar{a},\bar{Z}_0} - Y_{0,\bar{Z}_0} \mid W) \\ &= E_W E_{\bar{Z}_0|W} E(Y_{\bar{a},\bar{Z}_0} - Y_{0,\bar{Z}_0} \mid \bar{Z}_0, W) \\ &= E_W \int E(Y_{\bar{a},\bar{z}} - Y_{0,\bar{z}} \mid \bar{Z}_0 = \bar{z}, W) dF_{Z_0|W}(\bar{z}) \end{aligned}$$

$$= \quad E_W \int E(Y_{\bar{a},\bar{z}} - Y_{0,\bar{z}} \mid W) dF_{Z_0|W}(\bar{z}) \text{ by (4)}$$

$\equiv \widetilde{\text{DE}}$, (6)where the right-hand side is identifiable from the observed data distribution.

We will present this identifiability result as a theorem.

Theorem 1 *Let $DE(\bar{a}) = E(Y_{\bar{a},\bar{Z}_0} - Y_{0,\bar{Z}_0})$. Assume that (4) holds. Then,*

$$\begin{aligned} DE(\bar{a}) &= \widetilde{DE}(\bar{a}) \\ &\equiv E_W \int \{E(Y_{\bar{a},\bar{z}} \mid W) - E(Y_{0,\bar{z}} \mid W)\} \, dF_{Z_0|W}(\bar{z}). \end{aligned} \tag{7}$$

In subsection 3.2, we compare our identifiability result with the identifiability results of Robins (2003) and Pearl (2000).

3.2 Comparison with identifying assumptions of Robins and Pearl.

Comparison with Pearl (2001): Pearl (2001) shows (using the structural equation framework) that, if

$$Y_{\bar{a},\bar{z}} \perp \bar{Z}_0 \mid W \text{ for all } \bar{z}, \tag{8}$$

then

$$EY_{\bar{a},\bar{Z}_0} = E_W E(Y_{\bar{a},\bar{z}} \mid W) dF_{\bar{Z}_0|W}(\bar{z}).$$

This can be shown in precisely the same manner as we did in (6) above. Clearly, this assumption also implies $\text{DE} = \widetilde{\text{DE}}$ (see (7)). Thus Pearl (2001) identifiability mapping is the same as ours (7), but it was based on a different assumption.

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

$$Y_{\bar{a},\bar{z}} = m(\bar{a}, \bar{z}, W, e), \tag{9}$$

where e is a random variable which is conditionally independent of $\bar{Z}_0$, given W.

It is of interest to compare our assumption (4) with (8). Comparison of (5) with (9) helps one to understand that our assumption is less restrictive. That is, we assume $Y_{\bar{a},\bar{z}} - Y_{0,\bar{z}}$ is a function of $(\bar{a}, \bar{z}, W, e)$, while (8) assumes

that $Y_{\bar{a},\bar{z}}$ is a function of $(\bar{a}, \bar{z}, W, e)$, for some random variable e conditionally independent of $\bar{Z}_0$, given W. Stated in words, Pearl (2001) 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 no treatment. In contrast, our assumption states that, within subgroups defined by baseline covariates, the *magnitude of individual direct effects* at a fixed level of the intermediate variable is a deterministic function of treatment, the level of the intermediate variable, and an exogenous error. Under our assumption, at a fixed level of z, an individual's counterfactual outcome under a given treatment, $Y_{\bar{a},\bar{z}}$, can depend on the individual's counterfactual outcome under no treatment, $Y_{\bar{0},\bar{z}}$. Generally, $Y_{0,\bar{z}}$ explains a lot of the variation in $Y_{\bar{a},\bar{z}}$, suggesting that our assumption is more reasonable.

Suppose that assumption (8) holds at two treatment values $\bar{a}$ and 0. In that case, we have that both counterfactual outcomes $Y_{\bar{a},\bar{z}}$ and $Y_{0,\bar{z}}$ are conditionally independent of Z_0, given W. One would now expect that the difference $Y_{\bar{a},\bar{z}} - Y_{0,\bar{z}}$ is also conditionally independent of $\bar{Z}_0$, given W: in fact, mathematically it follows that $Y_{\bar{a},\bar{z}} - Y_{0,\bar{z}}$ is uncorrelated with any real valued function of $\bar{Z}_0$, given W. This suggests that in most examples in which (8) 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 (8) fails to hold.

Returning to our HIV example, consider the case that $\bar{Z}_0$ is the viral load of an HIV-infected person under no treatment, and $Y_{\bar{a},\bar{z}}$ is the CD4 T-cell count measured at the end of the study under a particular treatment regime $\bar{a}$ and a controlled viral load $\bar{z}$. Current understanding of HIV treatment suggests that the subject's counterfactual CD4 T-cell count $Y_{\bar{a},\bar{z}}$ is unlikely to be a deterministic function of the treatment regime, viral load, measured baseline factors, and an exogenous error. That is, one suspects that subjects have very different baseline CD4 T-cell count $Y_{0,\bar{z}}$-values, which are themselves extremely predictive of the counterfactual CD4 T-cell count $Y_{\bar{a},\bar{z}}$

under treatment regime $\bar{a}$, 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 therapy is likely to depend on what that individual's CD4 T-cell count would have been under no therapy. In this case the assumption of Pearl (2001) does not hold. However, it seems less unreasonable to assume that, within subpopulations defined by baseline factors W, all this variation in $Y_{\bar{a},\bar{z}}$ is explained by $Y_{0,\bar{z}}$, in the sense

that $Y_{\bar{a},\bar{z}} - Y_{0,\bar{z}}$ is a deterministic function of $\bar{a}, \bar{z}, W$ and an "exogenous" error. In other words, within subpopulations defined by baseline covariates and fixed viral load z, the magnitude of the direct effect of antiretroviral therapy on an individual does not depend on what that individual's CD4 T-cell count would have been under no therapy.

Comparison with Robins (2003): We refer to Robins (2003) for further discussion of the limitations of assumption (8). Robins proposes an alternative identifying assumption, which he calls the No-Interaction Assumption:

$$Y_{\bar{a},\bar{z}} - Y_{0,\bar{z}} \text{ is a random function } B(\bar{a}) \text{ that does not depend on } \bar{z}. \quad (10)$$

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. Clearly, under this assumption we have $Y_{\bar{a},\bar{Z}_0} - Y_{0,\bar{Z}_0} = Y_{\bar{a},\bar{z}} - Y_{0,\bar{z}}$ for any $\bar{z}$ so that

$$E(Y_{\bar{a},\bar{Z}_0} - Y_{0,\bar{Z}_0}) = E(Y_{\bar{a},\bar{z}} - Y_{0,\bar{z}}), \quad (11)$$

where the latter quantity does not depend on z.

A detailed mechanistic discussion of this assumption is given in Robins and Greenland (1992). As noted by Pearl (2001), this assumption is satisfied in the usual linear SEM model and has been used to identify direct and indirect effects in the structural equation literature.

The "No-interaction Assumption" implies, in particular, that $EY_{\bar{a},\bar{z}} = m_1(\bar{a}) + m_2(\bar{z})$ 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. This assumption can be tested by testing for an interaction term in a marginal structural model. In most applications one expects these interactions to be present, and, in fact, the interactions themselves often correspond with interesting and important 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 antiretroviral treatment at a controlled viral load does not depend on the level at which viral load is controlled. In other words, it implies that the direct effect of 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, some antiretroviral

drugs are hypothesized to act directly on CD4 T-cells by inhibiting their apoptosis (programmed cell death) (Phenix et al. (2000)). Higher levels of ongoing CD4 T-cell apoptosis may be induced by higher viral loads (Muthumani et al. (2003)). Thus, we could hypothesize that, if therapy has an anti-apoptotic direct effect on CD4 T-cell count (ie, not mediated by changes in viral load), such an effect may larger among individuals with higher viral loads and higher levels of apoptosis. In such a case, Robins' assumption does not hold. In contrast, our assumption simply requires that the magnitude of the individual direct effect at a fixed viral load be independent of what the individual's viral load would have been under no treatment.

Robins (2003)' identifiability mapping (11) corresponds with ours using an empty W (and thus with Pearl (2000)'s), since the integration w.r.t. $F_{\bar{Z}_0}$ does not affect the integral. We conclude that all three identifiability mappings agree with each other (except that Robins (2003) 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 (2001) 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 (8) assumption holds, or the "No-Interaction Assumption" holds.

It is interesting to note that our assumption corresponds with making the assumption (8) as in Pearl (2001), but replacing $Y_{\bar{a},\bar{z}}$ by the difference $Y_{\bar{a},\bar{z}} - Y_{0,\bar{z}}$. Thus our assumption can be viewed as a combination of the ideas presented in the two assumptions (8) and (10).

3.3 Discussion of assumption in terms of structural equation models.

In this subsection we will provide examples of structural equation models in which our assumption (4) holds. For example, consider the following semiparametric structural equation model for the data generating mechanism of the simplest single time-point version $O = (W, A, Z, Y)$ of our longitudinal data structure:

$$\begin{aligned} W &= f_1(U, e_1),\ e_1 \text{ exogenous},\ U \text{ unobserved characteristics} \\ A &= f_2(W, e_2),\ e_2 \text{ exogenous} \\ Z &= f_3(U, A, W, e_3),\ e_3 \text{ exogenous} \end{aligned}$$

$$Y \;=\; g(A, Z, W, e_{41}) + f_4(U, W, Z, e_{42}),\ e_{41}, e_{42} \text{ exogenous.} \qquad (12)$$

In this model the functions $f_1, f_2, f_3, f_4,\ g$ are arbitrary, and U can also follow an arbitrary distribution. Note that, beyond the assumption that A is randomized conditional on W (second equation), the main assumption of this model is that the effect of A on Y is additive w.r.t. to U (last equation). That is, this model does not allow an interaction between A and U, but it does allow an interaction of A with Z and W. Under these assumptions we have that $Y_{a,z} - Y_{0,z} = g(a, z, W, e_{41}) + f_4(U, W, z, e_{42}) - g(0, z, W, e_{41}) - f_4(U, W, z, e_{42}) = g(a, z, W, e_{41}) - g(0, z, W, e_{41})$, and $Z_0 = f_3(U, 0, W, e_3)$. Thus, if e_{41} is independent of e_3, given W, or $g(a, z, W, e_{41}) - g(0, z, W, e_{41})$ does not depend on e_{41} (that is, the error is additive), then assumption (4) holds. On the other hand, assumption (8) fails in this case. In the above structural equation model Z is confounded by unmeasured confounders. If we do not allow this, then (8) also holds.

By introducing a variable $L(1)$ (affected by past $(A(0), Z(0))$) between a $Z(0)$ and $Z(1)$, it can be shown that (4) holds under similar non-interaction constraints, but now also on the equation for $L(1)$. In this case, the assumption (8) fails to hold even when $(A(j), Z(j))$ is sequentially randomized. Consider the following structural equation model

$$\begin{aligned}
W &= f_0(U, e) \\
A(0) &= f_1(e) \\
Z(0) &= f_2(W, A(0), e) \\
L(1) &= f_3(W, A(0), Z(0), U, e) \\
A(1) &= f_4(W, A(0), Z(0), L(1), e) \\
Z(1) &= f_5(W, A(0), A(1), Z(0), L(1), e) \\
Y &= f_6(W, A(0), A(1), Z(0), Z(1), L(1), U, e)
\end{aligned}$$

with the following constraints:

$L(1) \to f_6(W, A(0), A(1), Z(0), Z(1), L(1), U, e)$ is linear
$f_3(W, a(0), z(0), U, e) - f_3(W, 0, z(0), U, e)$ is a function of $(W, a(0), z(0), e)$ only
$f_6(W, \bar{a}, \bar{z}, L(1), U, e) - f_6(W, 0, \bar{z}, L(1), U, e)$ is a function of $(W, \bar{a}, \bar{z}, e)$ only.

Again, let e denote here an exogenous random variable. For example, $f_3(W, a(0), z(0), U, e) = f_{31}(W, a(0), z(0), e) + f_{32}(W, z(0), U, e)$ (i.e., no-interaction between A and U),and $f_6(W, \bar{a}, \bar{z}, L(1), U, e) = f_{61}(W, \bar{a}, \bar{z}, e) + f_{62}(W, \bar{z}, L(1), U, e)$ (i.e., no-interaction between A and $(L(1), U)$). In the same manner as above one can

verify that our assumption (4) holds. On the other hand, if $L(1)$ is a function of U, then no restrictions on the functions f_3 and f_6 validate assumption (8).

Stated in Directed Acyclic Graph terminology, Pearl (2000) and Robins (2003) point out that assumption (8) will hold if, and essentially only if, there is no descendant of A that is also an ancestor of both Z and Y. In contrast, our assumption (4) holds in many cases where this condition is not met. We use a simple example, represented in Figure 2, to illustrate our point, relying again on antiretroviral treatment of HIV.

Consider the following structural equation model

$$\begin{aligned} A &= f_1(e_1) \\ R &= f_2(A, e_2) \\ Z &= f_3(A, R, e_3) \\ Y &= f_4(A, Z, R, e_4) = f_{41}(A, Z, e_{41}) + f_{42}(Z, R, e_{42}) \end{aligned}$$

Let A denote antiretroviral treatment, R denote viral resistance to treatment, Z denote viral load, and Y denote CD4 T-cell count at the end of the study. Again, let e_i denote an exogenous random variable. Note that the model does not allow an interaction between A and R, but it does allow an interaction of A with Z and Z with R. In other words, the effects of antiretroviral therapy and viral resistance on CD4 T-cell count are additive at a given viral load, but they can differ depending on the level of viral load. Under these assumptions, we have that $Y_{a,z} - Y_{0,z} = f_{41}(a, z, e_{41}) + f_{42}(z, R, e_{42}) - f_{41}(0, z, e_{41}) - f_{42}(z, R, e_{42}) = f_{41}(a, z, e_{41}) - f_{41}(0, z, e_{41})$ and $Z_0 = f_3(0, R, e_3)$. As above, if $f_{41}(a, z, e_{41}) - f_{41}(0, z, e_{41})$ does not depend on e_{41} (the error is additive), or if e_{41} is independent of e_3 (the error of the individual direct effect at a fixed viral load does not depend on the error of the viral load under no treatment), then our assumption holds. However, the assumption of Pearl (2000) fails, despite randomization of A and Z given the observed past.

For the sake of space, we will omit further discussion of the relation between structural equation models and our assumption (4).

4 Estimation of direct causal effects with MSM's.

Marginal structural models are models for marginal distributions of treatment-specific counterfactuals, possibly conditional on baseline covariates (e.g., Robins (1997), Robins (2000), van der Laan and Robins (2002)). These

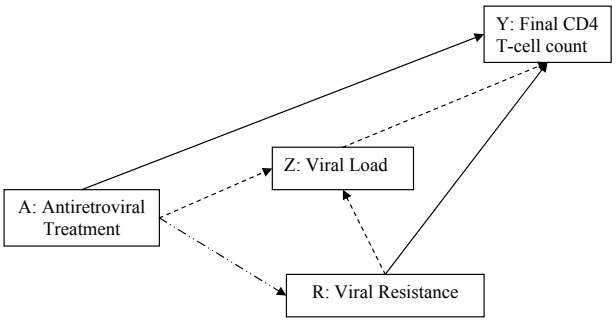

Figure 2:

models provide us with a natural approach to statistical inference regarding direct effects.

4.1 Test for no direct effect

Consider a marginal structural model $E(Y_{\bar{a},\bar{z}} \mid W) = g(\bar{z}, W \mid \alpha) + \gamma(\bar{a}, \bar{z}, W \mid \beta)$ indexed by parameters (α, β), where $\gamma(0, \bar{z}, W \mid \beta) = 0$ for all $\bar{z}, W, \beta$. Then $E(Y_{\bar{a},\bar{z}} - Y_{0,\bar{z}} \mid W) = \gamma(\bar{a}, \bar{z}, W \mid \beta)$. For example, if a is univariate, then one could pose the model

$$\gamma(a, z, W \mid \beta) = a * (\beta_0 + \beta_1 z + \beta_2 W).$$

Under the assumptions SRA, ETA, and (4), in such a parameterization we have that $\beta = 0$ implies that there is no direct effect. In addition, if $\beta \neq 0$, then only a miraculous cancellation would cause the direct effect to equal zero. Therefore, we believe that it is appropriate to test for no direct causal effect by testing $H_0 : \beta = 0$.

We can estimate the parameters of the marginal structural model with the IPTW, DR-IPTW or G-computation estimators as presented in van der Laan and Robins (2002). We recommend that the model $g_0(\bar{z}, W \mid \alpha)$ for $E(Y_{0,\bar{z}} \mid W)$ is made as flexible as sample size allows, possibly using cross-validation-based model selection methods developed in van der Laan, Dudoit (2003). This approach allows the model to be restrictive only on the parameter of interest, but not on the nuisance parameters. This provides us now with an estimator of β.

Given an estimate β_n of the true parameter value β_0, we can test the null hypothesis $H_0 : \beta_0 = 0$ with the test statistic $(\beta_n - 0)/\widehat{\mathrm{SE}}(\beta_n)$, which is asymptotically distributed as a $N(0, 1)$-random variable.

4.2 Estimation.

We now consider estimation of $\widetilde{\mathrm{DE}} = E_W \int \{E(Y_{\bar{a},\bar{z}} \mid W) - E(Y_{0,\bar{z}} \mid W)\} dF_{Z_0|W}(\bar{z})$. Let $\gamma(\bar{a}, \bar{z}, W \mid \beta)$ be a model for $E(Y_{\bar{a},\bar{z}} \mid W) - E(Y_{0,\bar{z}} \mid W)$ so that $\widetilde{\mathrm{DE}} = E_W E(\gamma(\bar{a}, \bar{Z}_0, W \mid \beta) \mid W)$. One can now pose a marginal structural model $E(Y_{\bar{a},\bar{z}} \mid W) = g_0(\bar{z}, W \mid \alpha) + \gamma(\bar{a}, \bar{z}, W \mid \beta)$. Above we discussed estimation of β. Alternatively, instead of first assuming a model for $E(Y_{\bar{a},\bar{z}} \mid W) - E(Y_{0,\bar{z}} \mid W)$, we could start out with posing a marginal structural model $E(Y_{\bar{a},\bar{z}} \mid W) = m(\bar{a}, \bar{z}, W \mid \lambda)$, and estimate its unknown

parameters as discussed above. This implies, in particular, a fit for $E(Y_{\bar{a},\bar{z}} \mid W) - E(Y_{0,\bar{z}} \mid W) = m(\bar{a}, \bar{z}, W \mid \lambda) - m(0, \bar{z}, W \mid \lambda)$. Given this estimate of $\gamma(\bar{a}, \bar{z}, W \mid \beta)$, it remains to estimate $E(\gamma(\bar{a}, \bar{Z}_0, W \mid \beta_n) \mid W)$, where β_n is treated as fixed in this conditional expectation.

Various approaches can be considered for estimation of this conditional expectation w.r.t. of $\bar{Z}_0$, given W. First, one could estimate the conditional distribution of $\bar{Z}_0$, given W, with the G-computation estimator, IPTW-estimator, or the DR-IPTW estimator according to a marginal structural model for the distribution $Z_{\bar{a}}$ and evaluate the fit of this marginal structural model at $\bar{a} = 0$. In the special case that $\gamma(\bar{a}, \bar{Z}_0, W \mid \beta)$ is linear in Z_0, we have

$$E(\gamma(\bar{a}, \bar{Z}_0, W \mid \beta) \mid W) = \gamma(\bar{a}, E(\bar{Z}_0 \mid W), W \mid \beta).$$

Thus, in this case it suffices to estimate $E(\bar{Z}_0 \mid W)$. However, even when $\gamma(\bar{a}, \bar{Z}_0, W \mid \beta)$ is non-linear, it is a sensible stategy to estimate each $\bar{a}$-specific univariate counterfactual expectation $E(\gamma(\bar{a}, \bar{Z}_0, W \mid \beta) \mid W)$ separately, by fitting a marginal structural model $E(\gamma(\bar{a}, Z_{\bar{a}*}, W \mid \beta) \mid W) = m(\bar{a}*, W \mid \gamma)$, and evaluating its fit at $\bar{a}* = 0$. Subsequently, one can smooth each of these $\bar{a}*$-specific estimates according to a model for $E(\gamma(\bar{a}, Z_0, W \mid \beta) \mid W)$.

4.3 A simple example.

As an example, consider the simple single time-point data structure W, A, Z, Y. For simplicity, we assume that all variables are univariate. In the single time-point case, one can use standard regression methods to test for and estimate a direct effect. Of course, in order to have that $\widetilde{DE} = DE$, we need to assume that within strata of W, (A, Z) is randomized (there is no confounding at the level of either treatment or the intermediate variable), and that the direct effects $Y_{az} - Y_{0z}$ at fixed z are independent of Z_0 (i.e., our assumption (4)).

Consider now $\widetilde{\mathrm{DE}}$. Under the assumption that (A, Z) is randomized w.r.t. W, we have $E(Y \mid A = a, Z = z, W) = E(Y_{az} \mid W)$ (this is the G-computation formula for $E(Y_{az} \mid W)$), 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

$$E(Y_{az} - Y_{0z} \mid W) = a(\beta_0 + \beta_1 z + \beta_2 W + \beta_3 zW). \tag{13}$$

Thus, one can use standard linear regression software to 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 (13) 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 $E(Z_a \mid W) = E(Z \mid A = a, W)$ (this is the G-computation formula for $E(Z_a \mid W)$), indexed by parameters λ. Thus, λ can be estimated with the least squares estimator for the regression of Z on A, W. By the linearity of (13) in z, it follows that the direct effect is now modeled as

$$DE = a\left\{\beta_0 + \beta_1 Em(0, W \mid \lambda) + \beta_2 EW + \beta_3 E\{m(0, W \mid \lambda)W\}\right\}.$$

An estimate of DE is obtained by replacing the regression parameters (β, λ) by their least squares estimators.

Applied to our HIV example, the expected CD4 T-cell count (Y) given treatment (A), viral load (Z), and baseline covariates $(E(Y \mid A, Z, W))$ is modeled as a linear regression, whose terms can be separated into two components: terms containing treatment, including all interactions between treatment and other covariates $(A(\beta_0 + \beta_1 Z + \beta_2 W + \beta_3 ZW))$, and terms not containing treatment $(\alpha_0 + \alpha_1 Z + \alpha_2 W + \alpha_3 ZW)$. We test for no direct effect by testing the hypothesis that the coefficients for the terms containing treatment in the multivariable regression model (the β's) are all equal to zero.

We now have a model for the direct effect at a fixed viral load z: $E(Y_{az} - Y_{0z} \mid W) = a(\beta_0 + \beta_1 z + \beta_2 W + \beta_3 zW)$. In order to estimate the direct effect, it remains to estimate viral load under no treatment: that is, $E(Z_0 \mid W) = E(Z \mid A = 0, W)$. Thus, we simply model the regression of viral load on treatment and baseline covariates, and evaluate our model fit under no treatment $(a = 0)$. The direct effect $E(Y_{aZ_0} - Y_{0Z_0} \mid W = w_i)$ for subject i with baseline covariate value w_i can now be estimated from the multivariable regression model of CD4 T-cell count by setting viral load at its expected value (i.e., $E(Z \mid A = 0, W = w_i)$) under no treatment and summing the coefficients of the terms containing treatment. Finally, the population direct effect DE is now estimated by taking the empirical mean of these subject-specific quantaties.

5 General direct causal effects

The previous sections have employed the definition of direct effect used by Robins and Greenland (1992) and Pearl (2000), where the direct effect is defined as the mean of individual counterfactual direct effects. In this section we present two generalizations of this definition, which we term "parameter of change" and "change of parameter", and extend our identifiability results to these general definitions.

5.1 Generalizing the definition of a direct effect.

Parameter of Change: A direct causal effect of $\bar{A}$ on Y can be generally defined as a parameter of the distribution of the individual direct effect $Y_{\bar{a},\bar{Z}_0} - Y_{0,\bar{Z}_0}$, where $\bar{Z}_0$ is the counterfactual of $\bar{Z}$ corresponding with setting $\bar{A} = 0$:

$$\text{DE} = \Phi(F_{Y_{\bar{a},Z_0} - Y_{0,Z_0}}).$$

For example, if one is concerned with estimation of the direct effect of treatment on the mean of the outcome of interest, then $\text{DE}(\bar{a}) = E(Y_{\bar{a},\bar{Z}_0} - Y_{0,\bar{Z}_0})$. The above definition of direct effect generalizes the definition of direct effect (1) on the mean of Y, as presented and considered in Robins and Greenland (1992), Robins (2003), and Pearl (2000).

Change of parameter: Alternatively, one can define a direct effect as a particular difference between the distribution $F_{Y_{\bar{a},\bar{Z}_0}}$ of $Y_{\bar{a},\bar{Z}_0}$ and the distribution $F_{Y_{0,\bar{Z}_0}}$ of $Y_{0,\bar{Z}_0}$. Let

$$\text{DE} = \Phi(F_{Y_{0,\bar{Z}_0}}, F_{Y_{\bar{a},\bar{Z}_0}})$$

represent such a difference.

For example, if Φ maps a cumulative distribution into its mean, then this reduces to $\text{DE}(\bar{a}) = EY_{\bar{a},\bar{Z}_0} - EY_{0,\bar{Z}_0}$ which agrees with (1). A more general real valued direct causal effect can be defined as

$$\text{DE}(\bar{a}) = \Phi(F_{Y_{0,\bar{Z}_0}}, F_{Y_{\bar{a},\bar{Z}_0}}) = \theta(F_{Y_{\bar{a},\bar{Z}_0}}) - \theta(F_{Y_{0,\bar{Z}_0}}),$$

where $\theta(F_{Y_{\bar{a},\bar{Z}_0}})$ denotes a real valued parameter of the distribution of $Y_{\bar{a},\bar{Z}_0}$. Alternatively, one could define the direct causal effect in terms of a quantile-quantile function:

$$\text{DE}(\bar{a}, y) = \Phi(F_{Y_{0,\bar{Z}_0}}, F_{Y_{\bar{a},\bar{Z}_0}})(y) \equiv y - F^{-1}_{Y_{\bar{a},\bar{Z}_0}} F_{Y_{0,\bar{Z}_0}}(y).$$

That is, the magnitude of the direct causal effect depends on the deviation of the quantile-quantile function from the identity function. We remind the reader that this quantile-quantile function maps a quantile (e.g., median) of $F_{Y_{0,\bar{z}_0}}$ into the same quantile of $F_{Y_{\bar{a},\bar{z}_0}}$.

5.2 Identification of direct effect as "parameter of change".

The approach presented in Section 4 for estimating direct causal effects in the mean sense under assumption (4) can be generalized to other definitions of direct causal effects in the following manner. Suppose we define a direct causal effect as a parameter of the distribution of $Y_{\bar{a},Z_0} - Y_{0,Z_0}$: $\mathrm{DE} = \Phi(F_{Y_{\bar{a},Z_0}-Y_{0,Z_0}})$. Under assumption (4) we have

$$F_{Y_{\bar{a},Z_0}-Y_{0,Z_0}}(\cdot) = E_W \int F_{Y_{\bar{a},\bar{z}}-Y_{0,\bar{z}}|W}(\cdot) dF_{Z_0|W}(\bar{z}).$$

We will state this result as a formal theorem, which generalizes Theorem 1.

Theorem 2 *Let* $DE = \Phi(F_{Y_{\bar{a},Z_0}-Y_{0,Z_0}})$. *Suppose that Assumption (4) holds. Then,*

$$\begin{aligned} DE &= \widetilde{DE} \\ &\equiv \Phi\left(E_W \int F_{Y_{\bar{a},\bar{z}}-Y_{0,\bar{z}}|W}(\cdot) dF_{Z_0|W}(\bar{z})\right) \end{aligned} \tag{14}$$

With the exception of the mean-case, we still need an additional model assumption (beyond SRA, ETA, and (4)) in order to identify $\widetilde{DE}$. Specifically, we need an assumption which allows one, within strata of W, to map the identifiable (by the G-computation formula) marginal distributions of $Y_{\bar{a},\bar{z}}$ and $Y_{0,\bar{z}}$ into the distribution of $Y_{\bar{a},\bar{z}} - Y_{0,\bar{z}}$. For example, we could assume that $Y_{\bar{a},\bar{z}}$ is a deterministic function of $Y_{0,\bar{z}}$, $\bar{a}$, $\bar{z}$, and W, which is equivalent with assuming that

$$Y_{0,\bar{z}} = F^{-1}_{Y_{0,\bar{z}}|W} F_{Y_{\bar{a},\bar{z}}|W}(Y_{\bar{a},\bar{z}}). \tag{15}$$

Under this assumption, within strata of W, the marginal distributions of $Y_{0,\bar{z}}$ and $Y_{\bar{a},\bar{z}}$ identify the joint distribution of $(Y_{0,\bar{z}}, Y_{\bar{a},\bar{z}})$, and thereby the distribution of $Y_{\bar{a},\bar{z}} - Y_{0,\bar{z}}$. Specifically,

$$F_{Y_{\bar{a},\bar{z}}-Y_{0,\bar{z}}}(\cdot) = F_{Y_{\bar{a},\bar{z}}} g^{-1}_{\bar{a},\bar{z},W}(\cdot),$$

where $g_{\bar{a},\bar{z},W}(x) \equiv x - F^{-1}_{Y_{0,\bar{z}}|W} F_{Y_{\bar{a},\bar{z}}|W}(x)$.

5.3 Identification of linear direct effects as "change of parameter".

Consider now the case that DE $= \Phi(F_{\bar{a},Z_0}, F_{0,Z_0}) = \theta(F_{\bar{a},Z_0}) - \theta(F_{0,Z_0})$, and that $\theta(F)$ is linear in F. Then we have:

$$\text{DE} \;=\; E_W \int \Phi(\left(F_{Y_{0,\bar{z}}|W,Z_0=\bar{z}}, F_{Y_{\bar{a},\bar{z}}|W,Z_0=\bar{z}}\right)) \, dF_{Z_0|W}(\bar{z}).$$

This proves the following identifiability result.

Theorem 3 *Let $DE = \Phi(F_{\bar{a},Z_0}, F_{0,Z_0}) = \theta(F_{\bar{a},Z_0}) - \theta(F_{0,Z_0})$, where $\theta(F)$ is linear in F. Consider the following assumption:*

$$\theta\left(F_{Y_{0,\bar{z}}|W,Z_0=\bar{z}}\right) - \theta\left(F_{Y_{\bar{a},\bar{z}}|W,Z_0=\bar{z}}\right) = \theta\left(F_{Y_{0,\bar{z}}|W}\right) - \theta\left(F_{Y_{\bar{a},\bar{z}}|W}\right). \qquad (16)$$

Then,

$$\begin{aligned} DE &= \widetilde{DE} \\ &\equiv E_W \int \theta\left(F_{Y_{0,\bar{z}}|W}\right) - \theta\left(F_{Y_{\bar{a},\bar{z}}|W}\right) dF_{Z_0|W}(\bar{z}) \end{aligned} \qquad (17)$$

Clearly, Assumption 16 holds under the assumption (8) of Pearl (2000) and Pearl (2001). The assumption (16) states in words that, within strata of W, the value of Z_0 should have an additive affect on the θ-parameter of the distribution of $Y_{\bar{a},\bar{z}}$ which does not depend on $\bar{a}$.

6 Discussion

Our results establish that the assumptions underlying the identifiability of direct and indirect causal effects are more realistic than those previously stated. As a consequence, statistical estimation of direct and indirect effects in longitudinal studies can be carried out within a reasonable statistical model. We plan to estimate (and test for) direct effects in a number of ongoing longitudinal studies, using marginal structural models and corresponding double robust inverse probability of treatment weighted estimators, as outlined in Section 4.

The statistical framework and results in this article immediately generalize to the following most generally defined longitudinal data structure (see

van der Laan and Robins (2002)), which allows one to define the outcome Y of interest as a survival time. Suppose that the observed data structure on a randomly sampled subject is defined as $O = (T, \bar{A}(T), \bar{Z}(T), \bar{L}(T))$, where T is a possibly random endpoint such as a survival time, $A(t)$ denotes treatment at time t, $Z(t)$ is the time-dependent intermediate covariate we control for in the definition of direct and indirect effects, and $L(t)$ denotes the remaining time-dependent measurements. Let $R(t) = I(T \leq t)$ so that T is identified by this time-dependent process $R(t)$. Suppose that the outcome Y of interest is a function of this data structure: $Y = f(\bar{R}(T), \bar{Z}(T), \bar{L}(T))$ for some function f. For all treatment regimes $\bar{a}$, we define the treatment specific end-point $T_{\bar{a}}$, and the truncated treatment specific process $Z_{\bar{a}}(t) = Z_{\bar{a}}(\min(t, T_{\bar{a}})$. For all joint regimes $(\bar{a}, \bar{z})$, we define the counterfactual end-points $T_{\bar{a},\bar{z}}$, $R_{\bar{a},\bar{z}}(t) \equiv I(T_{\bar{a},\bar{z}} \leq t)$, the truncated counterfactual L-process $L_{\bar{a},\bar{z}}(t) = L_{\bar{a},\bar{z}}(\min(t, T_{\bar{a},\bar{z}}))$, and let $Y_{\bar{a},\bar{z}} = f\left(T_{\bar{a},\bar{z}}, \bar{z}, \bar{L}_{\bar{a},\bar{z}}(T_{\bar{a},\bar{z}})\right)$ be the corresponding counterfactual outcome. Let $\bar{R}_{\bar{a},\bar{z}}$, $\bar{L}_{\bar{a},\bar{z}}$, and $\bar{Z}_{\bar{a}}$ denote the full sample paths of these counterfactual processes. One could define $\bar{L}_{\bar{a}} = \bar{L}_{\bar{a},\bar{Z}_{\bar{a}}}$, and $Y_{\bar{a}} = Y_{\bar{a},\bar{Z}_{\bar{a}}}$. The full data structure is now defined as the vector

$$X = (\bar{Z}_{\bar{a}}, \bar{L}_{\bar{a},\bar{z}}, \bar{R}_{\bar{a},\bar{z}} : \bar{a}, \bar{z}),$$

which thus includes $Y_{\bar{a},\bar{z}}$. The observed data structure can now be viewed as a missing data structure on X:

$$O = (\bar{A}, \bar{Z} \equiv \bar{Z}_{\bar{A}}, \bar{L}_{\bar{A},\bar{Z}}, \bar{R}_{\bar{A},\bar{Z}}).$$

All definitions of direct and indirect effect and corresponding identifiability results can now be applied. For example, one can define a direct effect $E(Y_{\bar{a},\bar{Z}_0} - Y_{0,\bar{Z}_0})$, and use the corresponding expressions.

References

S. Deeks, J. Barbour, Martin J., Swanson M., and 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*, 181(3):946–53, 2000.

R. Gill and J.M. Robins. Causal inference in complex longitudinal studies: continuous case. *Ann. Stat.*, 29(6), 2001.

K. Muthumani, A. Choo, D. Hwang, M. Chattergoon, N. Dayes, D. Zhang, M. Lee, U. Duvvuri, and D. Weiner. Mechanism of HIV-1 viral protein R-induced apoptosis. *Biochemical and Biophysical Research Communications*, 304:583–92, 2003.

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

J. Pearl. Cognitive systems laboratory. Technical report, University of California, Los Angeles, Department of Computer Science, 2001.

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

J.M. Robins. The analysis of randomized and non-randomized aids treatment trials using a new approach in causal inference in longitudinal studies. In L. Sechrest, H. Freeman, and A. Mulley, editors, *Health Service Methodology: A Focus on AIDS*, pages 113–159. U.S. Public Health Service, National Center for Health SErvices Research, Washington D.C., 1989.

J.M. Robins. Correcting for non-compliance in randomized trials using structural nested mean models. *Communications in Statistics*, 23:2379–2412, 1994.

J.M. Robins. Causal inference from complex longitudinal data. In Editor M. Berkane, editor, *Latent Variable Modeling and Applications to Causality*, pages 69–117. Springer Verlag, New York, 1997.

J.M. Robins. Robust estimation in sequentially ignorable missing data and causal inference models. In *Proceedings of the American Statistical Association*, 2000.

J.M. Robins. Semantics of causal dag models and the identification of direct and indirect effects. In N. Hjort P. Green and S. Richardson, editors, *Highly Structured Stochastic Systems*, pages 70–81. Oxford University Press, Oxford, 2003.

J.M. Robins and S. Greenland. Identifiability and exchangeability for direct and indirect effects. *Epidemiology*, 3(0):143–155, 1992.

M.J. van der Laan and J.M. Robins. *Unified methods for censored longitudinal data and causality*. Springer, New York, 2002.

Z. Yu and M.J. van der Laan. Construction of counterfactuals and the g-computation formula. Technical report, Division of Biostatistics, UC Berkeley, 2002.