Title

Assumption Trade-Offs When Choosing Identification Strategies for Pre-Post Treatment Effect Estimation: An Illustration of a Community-Based Intervention in Madagascar.

Journal

Journal of causal inference, 3(1)

ISSN

2193-3677

Authors

Weber, Ann M
van der Laan, Mark J
Petersen, Maya L

Publication Date

2015-03-01

DOI

10.1515/jci-2013-0019

Peer reviewed

HHS Public Access

Author manuscript

J Causal Inference. Author manuscript; available in PMC 2016 March 01.

Published in final edited form as:

J Causal Inference. 2015 March ; 3(1): 109–130. doi:10.1515/jci-2013-0019.

Assumption Trade-Offs When Choosing Identification Strategies for Pre-Post Treatment Effect Estimation: An Illustration of a Community-Based Intervention in Madagascar

Ann M. Weber,
Stanford University, Stanford, CA, USA

Mark J. van der Laan, and
University of California – Berkeley, Berkeley, CA, USA

Maya L. Petersen
University of California – Berkeley, Berkeley, CA, USA

Ann M. Weber: annweber@stanford.edu; Mark J. van der Laan: laan@berkeley.edu; Maya L. Petersen: mayaliv@berkeley.edu

Abstract

Failure (or success) in finding a statistically significant effect of a large-scale intervention may be due to choices made in the evaluation. To highlight the potential limitations and pitfalls of some common identification strategies used for estimating causal effects of community-level interventions, we apply a roadmap for causal inference to a pre-post evaluation of a national nutrition program in Madagascar. Selection into the program was non-random and strongly associated with the pre-treatment (lagged) outcome. Using structural causal models (SCM), directed acyclic graphs (DAGs) and simulated data, we illustrate that an estimand with the outcome defined as the post-treatment outcome controls for confounding by the lagged outcome but not by possible unmeasured confounders. Two separate differencing estimands (of the pre- and post-treatment outcome) have the potential to adjust for a certain type of unmeasured confounding, but introduce bias if the additional identification assumptions they rely on are not met. In order to illustrate the practical impact of choice between three common identification strategies and their corresponding estimands, we used observational data from the community nutrition program in Madagascar to estimate each of these three estimands. Specifically, we estimated the average treatment effect of the program on the community mean nutritional status of children 5 years and under and found that the estimate based on the post-treatment estimand was about a quarter of the magnitude of either of the differencing estimands (0.066 SD vs. 0.26–0.27 SD increase in mean weight-for-age z-score). Choice of estimand clearly has important implications for the interpretation of the success of the program to improve nutritional status of young children. A careful appraisal of the assumptions underlying the causal model is imperative before committing to a statistical model and progressing to estimation. However, knowledge about the data-generating process must be sufficient in order to choose the identification strategy that gets us closest to the truth.

Correspondence to: Ann M. Weber, annweber@stanford.edu.

Keywords

causal effect; average treatment effect; community-level intervention; difference-in-differences; change score

Introduction

Interventions scaled-up to a large or national level have failed to consistently demonstrate the causal benefits anticipated by results of small-scale experimental studies [1]. Challenges in evaluating a program operating at-scale are not limited to the logistical and technical constraints of surveying hundreds to thousands of households across a region or country, but also include the design and analysis of an evaluation to determine whether the program actually results in a benefit, and if so to quantify it. Our ability to make a causal claim from an evaluation may easily be compromised by the choices we make in the analysis and are particularly complex in observational studies. These choices are made even more controversial with the availability of pre-treatment outcome data, as we will illustrate in this paper. Misleading estimates of a program's benefit (in either direction) have significant policy and funding implications for the program, as well as for the people the program is intended to help.

In an introductory chapter on econometric evaluations of social programs, Nobel laureate James Heckman and co-author Edward Vytlacil point out that we often confuse the three main issues that face an evaluation: definition, identification and estimation. The authors state that “particular methods of estimation (e.g. matching or instrumental variable estimation) have become associated with ‘causal inference’ and even the definition of certain ‘causal parameters’ …” [2]. Investigators from different disciplines will bring distinct theoretical and analytical frameworks to estimation, which can lead to differing estimates of the causal effect and contradictory conclusions, in some cases without strong theoretical justification for the approach used. However, the selection of an estimator should happen after defining the causal target parameter and after the underlying assumptions necessary to identify the parameter are made explicit. In this paper, we work step-by-step through a roadmap for causal inference [3, 4], and emphasize the need to “define first, identify second, and estimate last” (quote from Judea Pearl's forward in *Targeted Learning* by van der Laan and Rose) [4, 5].

We use a national nutrition program in Madagascar for illustration. The program was implemented at the community level: it was made available to all residents of a community and sharing of information within a community was encouraged. Importantly, the Madagascar program rollout was non-random: communities were selected for treatment if they were located within districts with a pre-program prevalence of childhood underweight[1] that was above the national average or if they met certain logistical criteria (e.g. a local non-profit organization was available to supervise the program). Cross-sectional anthropometric

[1]Underweight is an indicator of being two standard deviations below the median weight of a reference population of well-nourished and healthy children of the same age and gender.

surveys were administered in the same communities in Madagascar pre- and post-program implementation, providing multiple options for identification of a causal effect.

In this paper, we define the outcome as either the post-treatment value or the change from pre- to post-treatment. We consider the long-standing controversy over the advantages and disadvantages of each, reviewed in Imai [6] and Maris [7]. Using these two outcomes, we identify three statistical parameters used for interventions with pre-post data that under different assumptions are equivalent to our causal target parameter of interest. We include two difference-in-differences models commonly used with pre-post data: a change score estimand and an outcome estimand where the outcome data from both time periods are combined, or "pooled" together (popular in the social sciences and econometrics literature) [6, 8–10]. We also include a conventional approach in which the pre-treatment outcome (or lagged outcome) is included in the conditioning set of covariates.

Our study adds to the existing literature on the trade-offs of these causal models by using semi-parametric structural equation models to avoid making assumptions about the underlying functional form of the data-generating distribution [5, 11]. In addition, we use graphical representations of these models (directed acyclic graphs or DAGs) to make the assumptions underlying our causal models transparent and understandable to contextual experts. We review how DAGs can be used for locating sources of dependencies among variables and show in a series of data simulations how the estimate of the target causal parameter diverges from the truth when the necessary assumptions for a given identifiability result fail to hold. Finally, we apply a semi-parametric efficient estimator (targeted maximum likelihood) for each of the three estimands to the observed data from Madagascar to estimate the average treatment effect (ATE) of the program. We demonstrate that the choice of causal model and corresponding identification strategy have important implications for conclusions regarding the success of the program at improving the nutritional status of young children.

Setting, data and notation

In Madagascar, approximately 30% of children under five are estimated to be underweight [12]. Underweight is a near-term marker for inadequate nutrition and is estimated to be responsible for the largest proportion of the death and disease burden associated with malnutrition [13]. In 1999, the Madagascar National Office of Nutrition (ONN) implemented a comprehensive community-level growth-monitoring and nutrition program, incorporating multiple activities that have been found to be associated with better child outcomes [14]. Prior to the roll-out of the national nutrition program in 1997/1998, a large-scale anthropometrics survey was performed in 420 communities in Madagascar. The survey was administered to a random sample of 14,148 households, 12,814 of which had children 5 years of age and under. Treatment assignment in 1999 was made at the community level, based primarily on district-level prevalence of moderate underweight among children 5 years of age and under (obtained from the 1997/1998 study). The program was phased in through 2002 and expanded to include new communities impacted by severe weather conditions in 2000 or impacted by political instability in 2002. In total, 3,600 project sites were reached. In 2004 a second anthropometric survey was administered to 10,704

households; 9,296 with children 5 years of age and under, in 446 participating and non-participating communities. We restricted our analytic sample to 410 of these communities; 26 were excluded that were not part of the baseline survey in 1997, and 10 communities from 6 urban districts were excluded because sites in these districts were opened in 2002 in response to a political crisis, such that the context of the intervention in these sites differed substantially from the remainder of the country.

The data in our analysis are from two cross-sectional surveys with different individuals included in each year (i.e. outcomes and covariates were measured on one set of subjects in 1997/1998 (time $t = 0$) and on another set of subjects in 2004 (time $t = 1$)). For ease of comparison of the different estimands, we set the community as the unit of analysis. The observed variables for our analysis included community-level covariates (e.g. population size); individual-level covariates aggregated to the community level as a mean or proportion (e.g. proportion of mothers with no education); and the nutrition program administered at the community level. Our outcome of interest is community mean weight-for-age for children under 5 years, one of the primary nutritional outcomes in children targeted by the program. Our observed data are described in Table 1.

Causal inference road map

We follow a road map that links our research question to inference, making the underlying assumptions explicit for the path between the two [3, 4]. First, we define precisely the research question. This may seem obvious, but is often not made clear. Second, we turn the research question and relevant background knowledge into a structural causal model (SCM) [5], which encodes information about the relationships between the variables with a series of non- or semi-parametric equations. Importantly, we assume that the SCM accurately represents the data-generating processes that gave rise to our observed data. This is the key link from counterfactual to observed data.

Given the SCM, we specify the causal parameter of interest in the third step.[2] The causal parameter is the parameter we would obtain under an ideal experiment and is defined using counterfactual notation. We use the phrasing “intervening to set the treatment” or “setting $A = a$” to refer to the hypothetical treatment conditions that we want to apply to the system when making causal contrasts. In this paper, we are interested in estimating the difference in the expectation of counterfactual outcomes if we were to intervene to set the treatment to 1 (receive treatment) versus to 0 (not receive treatment), for all communities. This contrast is known as the ATE.

In the fourth step, we assess identifiability, or whether the observed data, in combination with our assumptions about the data-generating system, are sufficient to express the target casual parameter of interest as a parameter of the distribution of the observed data alone. This second parameter is the statistical target parameter (also referred to as the estimand; we use the terms interchangeably). *The estimand is the parameter we estimate directly with the*

[2]Note that our objective in this paper is not to argue for one single type of causal model (SCM), nor to say that the causal model specification must come before specifying the causal parameter, but rather to emphasize separation of the steps, and to make clear what the assumptions are when a given SCM is used.

observed data; under additional causal assumptions, it is equal to the casual parameter. We evaluate three different statistical target parameters commonly used with pre-post data, each of which is equivalent to the causal parameter under alternative identifying assumptions.

In the last steps of the roadmap, we commit to an estimand and statistical model and proceed using targeted maximum likelihood estimation, an efficient double robust approach. We present results using the observed data from Madagascar. In addition, we use simulations to illustrate the different assumptions required for the three statistical parameters to be equivalent to the ATE, our target parameter, of interest and potential consequences when the assumptions do not hold.

We could evaluate other causal parameters of interest, such as the average treatment effect among the treated, or the ATT, which is a conventional target parameter in the field of econometrics and impact evaluations of observational studies [2] and is of interest in the public health field. We chose the ATE to demonstrate how selecting a statistical model without understanding the underlying assumptions can threaten the validity of a causal effect estimate. The choice of a different causal parameter would not eliminate this threat.

The research question, SCM and target causal parameter

Our causal question is: Does the intervention increase the average nutritional status of children living in the community? In this paper, we are interested in estimating a population average effect at the community level, for all communities in the target population.

The SCM is characterized by a set of endogenous variables at two time points (see notation Table 1). Community variables that are not aggregates of individual factors are denoted by V and are assumed to be time invariant for the period of the study. Individual-level factors aggregated up to community-level factors are denoted by a vector, $W^c(t)$, at time $t = 0,1$. The community-level mean outcome for children at time t is denoted as $Y^c(t)$. The community-level exposure, A is assigned to zero or one as a function of V, $W^c(t = 0)$ and $Y^c(t = 0)$. In addition, there are unmeasured exogenous variables, U, that may cause random variation in each of the observed variables. Restrictions on the joint distribution of these unmeasured errors will be required for identifiability.

We pose the following SCM to explain the relationships between the variables:

$$
\begin{aligned}
V &= f_V(U_V)\\
W^c(t=0) &= f_{W(t=0)}(V, U_{W(t=0)})\\
Y^c(t=0) &= f_{Y(t=0)}(V, W^c(t=0), U_{Y(t=0)})\\
A &= f_A(V, W^c(t=0), Y^c(t=0), U_A)\\
W^c(t=1) &= f_{W(t=1)}(V, W^c(t=0), Y^c(t=0), U_{W(t=1)})\\
Y^c(t=1) &= f_{Y(t=1)}(V, W^c(t=0), Y^c(t=0), A, W^c(t=1), U_{Y(t=1)}),
\end{aligned}
$$

where no assumptions are made about the form of the functions. Our target causal parameter is the ATE given by: $E(Y_1^c(t=1) - Y_0^c(t=1))$, where $Y_a^c(t)$ denotes the counterfactual community-level outcome under an intervention on the SCM setting $A = a$.

We start with a model with a minimal set of exclusion restriction assumptions about the data-generating system in order to avoid imposing restrictions that may or may not be reflective of the true data-generating process. We make a single exclusion restriction in this model: that the covariates $W^c(t = 1)$ occurring post intervention are not affected by the intervention. We impose this exclusion restriction for three reasons. First, it is a reasonable assumption in the context of the Madagascar study. Second, it is required for one of the three estimands (see identifiability section for estimand III), and we apply it to the other two to facilitate our comparison across estimands. Finally, it allows us to condition on $W^c(t = 1)$ in the models to better predict $Y^c(t = 1)$. Although we cannot test whether this exclusion restriction holds, we find that A is not significantly associated with any of the variables in $W^c(t = 1)$, when controlling for the corresponding variable in $W^c(t=0)$ using a series of parametric regressions (data not shown). In addition, our estimation results were not found to be sensitive to removing $W^c(t = 1)$ from the set of conditioning variables for estimands I and II (data not shown).

Identifiability

Causal effect estimation relies on assumptions that must be made explicit when using observational data for causal inference. Specifically, some form of the randomization assumption (RA) and the experimental treatment assignment (ETA) assumption are sufficient for our causal parameter to be identified.[3]

The RA (also known as the assumption of no unmeasured confounders, or of exchangeability), states that treatment, A, is independent of counterfactual outcome, Y_a, given some subset of the data. The RA is a causal assumption, and as such is not testable. However, we can draw a graphical representation of our SCM (i.e. a DAG) to check whether our assumptions about the underlying data-generating system are sufficient to imply that our identifying assumptions hold [5, 11]. By using a graphical procedure, we are able to solve the identification problem without resorting to an algebraic analysis of whether a statistical model parameter has a unique solution in terms of the parameters of the distribution of the observed variables [5, 18]. Very briefly, the graph is drawn based on the relationships defined in the SCM, where the parents of a variable (variables on the right-hand side of the equation) are connected to the child variable (variable on the left-hand side of the equation) with an arrow directed toward it. A path is any sequence of lines connecting two variables. The arrow between two variables can only go in one direction, such that the paths are acyclic (i.e. the graph cannot have $A \rightarrow B \rightarrow C \rightarrow A$). Paths can either be open or blocked, depending on the direction of the arrows and whether or not a variable is conditioned on. (In this paper, we represent conditioning on a variable by placing a box around it.) Open paths can give rise to dependency between variables, and the absence of any open paths implies independence.

[3]The consistency assumption and the stable unit treatment value assumption (SUTVA) typically associated with the Rubin framework [15, 16] are subsumed in our SCM. The consistency assumption states that an individual's (or community's) potential outcome under the treatment actually received is precisely the observed outcome. Our SCM already implies the counterfactual and provides the necessary link to the observed data. In addition, the SCM assumes that the data-generating system for each community is generated independently of the others, such that the absence of hierarchical relationships between communities implies that one community's outcome is unaffected by another's treatment assignment (i.e. SUTVA holds) [17].

The specific RA and necessary additional assumptions for our three estimands are discussed in detail below. To minimize confusion from too many arrows, we represent DAGs for each estimand using a simplified data structure that omits the observed, time invariant, village factors, V. We justify this simplification because V are exogenous to the data-generating system (no arrows go into V, other than U_V) and if we condition on V, we do not have to worry about unblocked paths from unmeasured variables through V. In most cases, we also omit the exogenous variables, U. The omission of the U's implies that these exogenous variables are independent (discussed further with Figure 1). Paths depicted in red in the figures represent unblocked paths between the treatment and outcome variables. In our DAGs, we make use of a dashed line to represent the association between two variables created by these unblocked paths. Specifically, a dashed line indicates that the path is open due to conditioning on a collider (two arrows go into the same variable).

An assumption of sufficient support in the observed data distribution is also required for the target statistical parameter to be identified. The strong ETA assumption (also known as the positivity assumption) states that there must be sufficient variation in treatment (i.e. some positive probability of both being treated and not being treated) within strata of confounders, although this can be weakened under additional assumptions.

In the Madagascar example, the ETA assumption was not theoretically violated. Communities in non-targeted districts participated and communities in targeted districts did not. Although most (92%) of communities in the non-targeted districts did not take up the program, only 66% of communities in targeted districts in our sample took up the program by 2004. In addition, many non-participating communities received the program after 2004 as the program expanded. As expected, the treated communities have on average a higher prevalence of underweight (39% vs. 30%), but there was reasonable heterogeneity in underweight by treatment status: the range among the treated communities was 5–95% prevalence, and 0–70% prevalence among the untreated communities.

However, the sample was finite and the covariate data were high dimensional, such that the ETA assumption may be practically violated. Due to the impossibility of checking every level of the covariates, we examined the distribution of estimated probabilities of treatment given our covariates and found that the probabilities are bounded between 0.025 and 0.975 in our sample for each of the estimands (data not shown). The range of probabilities in the untreated group is comparable to that of the treated group. Although these checks do not quantify the degree to which violations or near-violations threaten the validity of our causal effect estimate, evidence of heterogeneity in treatment within strata of the confounders gave us some confidence that the ETA assumption is reasonably held. A formal diagnostic based on the parametric bootstrap is available for estimating the presence and magnitude of bias from positivity violations and near-violations [19, 20], but was not performed here.

In the next sections, we describe three identifiability results (and corresponding estimands) where we link the causal parameter to our observed data distribution. In the first estimand, the outcome is defined as the outcome post treatment, $Y^c(t=1)$. In the second estimand, the outcome is defined as the change in outcome pre- vs. post-treatment, Y^{θ}; and in the third, the

outcome combines the data from both time periods, $Y^c(t)$. We refer to the latter as the pooled outcome estimand.

Estimand I: outcome $Y^c(t=1)$

For the first estimand with outcome $Y^c(t = 1)$, identifiability is based on conditioning on all baseline covariates, including the pre-treatment (or lagged) outcome, as well as post-treatment covariates W^c (t=1) assumed not to be affected by A, as discussed above. The RA for this estimand is:

$$Y^c_a(t=1) \perp A | V, W^c(t=0), Y^c(t=0), W^c(t=1) \quad (1)$$

For the RA(1) to hold, it is sufficient that the exogenous variables for the exposure, U_A, be independent of the exogenous variables for the outcome, $U_{Y(t=1)}$, given V, $W^c(t = 0)$, $Y^c(t = 0)$, $W^c(t = 1)$. This additional independence assumption is reasonable if we have no unmeasured common causes of A and $Y^c(t= 1)$ (i.e. no confounders).

The DAG in Figure 1 encodes the information from the series of equations in the SCM and allows us to visually check that $Y^c_a(t=1)$ is independent of A given $W^c(t = 0)$, $Y^c(t=0)$, and $W^c(t =1)$. Specifically, we verify that our conditioning variables, which must not be affected by the intervention, block any unblocked backdoor path from A to $Y^c(t =1)$, while not opening any new paths. This is referred to as satisfying the backdoor criterion [18]. The RA(1) holds under this model.

We now have the following identifiability result:

$$\begin{aligned} E(Y^c_a(t=1)|V, W^c(t &=0), Y^c(t=0), W^c(t=1)) \\ &= E(Y^c_a(t=1)|A, V, W^c(t=0), Y^c(t=0), W^c(t=1)) \\ &= E(Y^c(t=1)|A, V, W^c(t=0), Y^c(t=0), W^c(t=1)) \end{aligned}$$

where the first equality holds under the RA(1), and the second holds under our definition of the counter-factual outcomes. Note that for these conditional expectations of the outcome to

be well defined without parametric model assumptions beyond those implied by our SCM, we need some communities with and without the treatment for each level of the conditioning variables V and $W^c(t)$ (i.e. we need for the ETA assumption to hold).

A first estimand (or statistical parameter) for the ATE, Ψ^I, follows:

$$\Psi^I(P_0)=E_{V,W(t=0),Y(t=0),W(t=1)}\left(\begin{array}{c}E(Y^c(t=1)|A=1,V,W^c(t=0),Y^c(t=0),W^c(t=1))\\-E(Y^c(t=1)|A=0,V,W^c(t=0),Y^c(t=0),W^c(t=1))\end{array}\right) \quad (2)$$

We refer to this estimand as the post-treatment estimand.

Estimand II: outcome Y^θ

Next, we consider the outcome as the change in the community specific means, Y^θ, before and after treatment. We define Y^θ as:

$$Y^\theta=Y^c(t=1)-Y^c(t=0) \quad (3)$$

By definition of the structural equations for $Y^c(t=1)$ and $Y^c(t=0)$, we have the following structural equation for Y^θ:

$$Y^\theta=f_{Y(t=1)}(V,W^c(t=0),Y^c(t=0),A,W^c(t=1),U_{Y(t=1)})-f_{Y(t=0)}(V,W^c(t=0),U_{Y(t=0)})$$

The DAG in Figure 2 reflects this same information. Note that $U_{Y(t=0)}$ now affects both $Y^c(t=0)$ and Y^θ, so we have included it in the graph. Under this model, we have a new RA for outcome, Y^θ:

$$Y^\theta_a \perp A|V,W^c(t=0),Y^c(t=0),W^c(t=1) \quad (4)$$

and we can identify a statistical target parameter based on Y^θ that is equivalent to Ψ^I [21]. Specifically, if we define the counterfactual mean of Y^θ_a under an intervention on the SCM setting $A = a$ as:

$$E(Y^\theta_a)=E\left(Y^c_a(t=1)-Y^c_a(t=0)\right)=E\left(Y^c_a(t=1)\right)-E\left(Y^c_a(t=0)\right) \quad (5)$$

then we can rewrite our target causal parameter in terms of Y^θ_a and show that it is identical to the ATE as previously defined as $E(Y^c_1(t=1)-Y^c_0(t=1))$. First, the parameter is expressed as a difference in the differences of means:

$$E(Y^\theta_1-Y^\theta_0)=(E(Y^c_1(t=1))-E(Y^c_1(t=0)))-(E(Y^c_0(t=1))-E(Y^c_0(t=0))) \quad (6)$$

However, since intervening to set the treatment cannot affect the pre-treatment outcome $(Y^{c}_{a}(t{=}0){=}Y^{c}(t{=}0))$, the above can be rewritten such that the mean of $Y^{c}(t{=}0)$ cancels out to give the ATE:

$$(E(Y^{c}{}_{1}(t{=}1)){-}E(Y^{c}(t{=}0))){-}(E(Y^{c}{}_{0}(t{=}1)){-}E(Y^{c}(t{=}0))){=}E(Y^{c}{}_{1}(t{=}1){-}Y^{c}{}_{0}(t{=}1)) \quad (7)$$

Under the RA(4), we can identify our statistical target parameter

$$\begin{aligned} E(Y^{\theta}_{a}|V, W^{c}(t &{=}0), Y^{c}(t {=}0), W^{c}(t {=}1)) \\ &{=}E(Y^{\theta}_{a}|A {=}a, V, W^{c}(t {=}0), Y^{c}(t{=}0), W^{c}(t {=}1)) \\ &{=}E(Y^{\theta}|A {=}a, V, W^{c}(t {=}0), Y^{c}(t {=}0), W^{c}(t{=}1)) \end{aligned}$$

and have an alternative, but equivalent, formulation of estimand Ψ^{I}:

$$\Psi^{I*}(P_0){=}E_{V,W(t=0),Y(t=0),W(t=1)}\begin{pmatrix} E(Y^{\theta}|A{=}1, V, W^{C}(t{=}0), Y^{C}(t{=}0), W^{C}(t{=}1)) \\ -E(Y^{\theta}|A{=}0, V, W^{C}(t{=}0), Y^{C}(t{=}0), W^{C}(t{=}1)) \end{pmatrix} \quad (8)$$

So what is the advantage of using Y^{θ} over $Y^{c}(t = 1)$ for estimating the ATE? The main justification in the causal inference literature is that the difference method allows for both the treatment, A, and outcome, $Y^{c}(t)$, to depend on unobserved community fixed effects that are time invariant [6, 22]. To explore this advantage, we add an unmeasured confounder, $C{=}f_{C}(U_{C})$, to our SCM and DAG, such that C is a common cause for A, $Y^{c}(t = 0)$, and $Y^{c}(t{=}1)$ (see Figure 3). The allowed functional forms of $f_{Y(t=0)}$ and $f_{Y(t=1)}$ in the SCM are restricted such that C has a linear additive effect on $Y^{c}(t)$, specifically that:

$$\begin{aligned} Y^{c}(t{=}0){=}f_{Y(t=0)}(V, W^{c}(t{=}0), U_{Y(t=0)}){+}C \\ Y^{c}(t{=}1){=}f_{Y(t=1)}(V, W^{c}(t{=}0), Y^{c}(t{=}0), A, W^{c}(t{=}1), U_{Y(t=1)}){+}C \end{aligned}$$

The introduction of an unmeasured confounder, C, opens up a backdoor path from A to $Y^{c}(t = 1)$ (see path $A \leftarrow C \rightarrow Y^{c}(t = 1)$ labeled (i) and colored red in Figure 3). The RA(1) for estimand I no longer holds. At first, it appears that RA(4) might hold for Y^{θ}. If we assume C has a constant additive effect on both $Y^{c}(t = 0)$ and $Y^{c}(t = 1)$, then Y^{θ} is not a function of C

when taking the difference of Y^c at the two time points. The structural equation for Y^θ remains unchanged in this case. Using Y^θ instead of $Y^c(t=1)$ as outcome has the potential (under this specific parametric assumption) to close one backdoor pathway from A to Y^θ via unmeasured confounder C.

However, on closer inspection, RA(4) does not hold under this model. Under the causal model where C affects $Y^c(t=0)$, $Y^c(t=1)$, and A, conditioning on $Y^c(t=0)$ induces new dependence between Y^θ and A and opens a backdoor path through exogenous variable $U_{Y(t=0)}$ and confounder C. This occurs because $Y^c(t=0)$ is a collider. Conditioning on a collider opens a path that would otherwise be blocked [11]. This unblocked path, $A \leftarrow C - U_{Y(t=0)} \rightarrow Y^\theta$, is represented by the dashed line between $U_{Y(t=0)}$ and C (labeled (ii) in Figure 4).

Thus, to benefit from the potential to remove unmeasured confounding from the use of Y^θ as outcome, we need a new RA(9), which is not conditional on $Y^c(t=0)$:

$$Y^\theta_a \perp A | V, W^c(t=0), W^c(t=1) \quad (9)$$

It is important to note that we have arrived at the same conclusion with DAGs that others have reached using parametric equations and analysis of covariance. In the econometrics literature, the problem is recognized as the fact that the residual on Y^θ (in a parametric equation) is necessarily correlated with the lagged outcome, $Y^c(t=0)$, because both are a function of the random error on $Y^c(t=0)$ (i.e. a function of $U_{Y(t=0)}$ in our SCM) [23]. Conditioning on $Y^c(t=0)$ has been demonstrated to bias the treatment effect estimate under this model where the errors on Y^c are serially correlated [23]. The method of differencing can still be applied if this correlation is thought to be negligible (e.g. possibly when the data are from a series of cross-sections of different individuals and/or the time between cross-sections is long) [24].

However, RA(9) still does not hold under this differencing model without additional assumptions. We make these assumptions apparent with the use of the DAG shown in Figure 5.

By not conditioning on $Y^c(t=0)$, we open up multiple new pathways from A to Y^θ: directly through $Y^c(t=0)$ ($A \leftarrow Y^c(t=0) \rightarrow Y^\theta$, labeled (iii) in Figure 5); through C ($A \leftarrow C \rightarrow Y^c(t=0) \rightarrow Y^\theta$, labeled (iv)); and through $U_{Y(t=0)}$ ($A \leftarrow Y^c(t=0) \leftarrow U_{Y(t=0)} \rightarrow Y^\theta$, labeled (v)). Additionally, $W^c(t=1)$ is a descendant of collider $Y^c(t=0)$, and conditioning on $W^c(t=1)$ opens up the same pathway as conditioning on $Y^c(t=0)$ (i.e. $A \leftarrow C - U_{Y(t=0)} \rightarrow Y^\theta$). However, if we do not condition on $W^c(t=1)$, then we would open up new backdoor pathways through $W^c(t=1)$ (i.e. $A \leftarrow C \rightarrow Y^c(t=0) \rightarrow W^c(t=1) \rightarrow Y^\theta$ and $A \leftarrow Y^c(t=0) \rightarrow W^c(t=1) \rightarrow Y^\theta$ labeled (vi)).

Therefore, we must be willing to make additional assumptions for our casual parameter to be identifiable in a difference model. Three additional exclusion restrictions are sufficient: $Y^c(t=0)$ must not affect A, $W^c(t=1)$ and $Y^c(t=1)$. We modify the structural equation for $Y^c(t=1)$ to

be an additive function of $Y^c(t=0)$, such that Y^θ no longer depends on $Y^c(t=0)$, and the semi-parametric equation for Y^θ becomes:

$$Y^\theta = f_{Y(t=1)}(V, W^c(t=0), A, W^c(t=1), U_{Y(t=1)}) + Y^c(t=0) - Y^c(t=0) = f_{Y\theta}(V, W^c(t=0), A, W^c(t=1), U_{Y(t=1)})$$

Under this model (see Figure 6), we can choose to either adjust for $W^c(t = 1)$ or not; conditioning on $W^c(t = 0)$ is sufficient and $W^c(t = 1)$ is no longer a descendant of a collider.

In summary, RA(9) holds in the presence of unmeasured confounding from non-time-varying factors, C, with a constant additive effect on $Y^c(t)$ if $Y^c(t=0)$ does not affect A, $Y^c(t = 1)$ and $W^c(t =1)$. The target causal parameter can now be identified as a new target parameter of the observed data distribution. The identifiability result applied to Y^θ becomes:

$$E(Y_a^\theta | V, W^c(t=0), W^c(t=1)) = E(Y_a^\theta | A=a, V, W^c(t=0), W^c(t=1)) = E(Y^\theta | A=a, V, W^c(t=0), W^c(t=1))$$

where the first equality holds under the RA(9) and the second from the definition of the counterfactual outcome Y^θ under our new SCM (Figure 6), giving us a new estimand for the ATE, Ψ^{II}:

$$\Psi^{II}(P_0) = E_{V,W(t=0),W(t=1)} \begin{pmatrix} E(Y^\theta | A=1, V, W^C(t=0), W^C(t=1)) \\ -E(Y^\theta | A=0, V, W^C(t=0), W^C(t=1)) \end{pmatrix} \quad (10)$$

Which we refer to as the change score estimand.

Estimand III: outcome $Y^c(t)$

Finally, there is an alternate difference-in-differences estimand that pools the outcome data from both time periods together. For this approach, we need to evaluate a third causal model for identifiability. Specifically, if we are willing to make additional assumptions on the underlying causal model such that:

$$E_{V,W(t=1),W(t=0)}(Y^c(t) | A=a, V, W^c(t=0), W^c(t=1)) = E_{V,W(t)}(Y^c(t) | A=a, V, W^c(t)), \text{ for } t=0,1 \quad (11)$$

then we have the following identifiability result under the new SCM:

$$
\begin{aligned}
&E(Y_{\mathrm{a}}^{\theta}|V,W^{\mathrm{c}}(t\\
&\quad=0),W^{\mathrm{c}}(t\\
&\quad\quad=1))\\
&=E(Y_{\mathrm{a}}^{\theta}|A\\
&=a,V,W^{\mathrm{c}}(t\\
&=0),W^{\mathrm{c}}(t\\
&\quad=1))\\
&=E(Y^{\mathrm{c}}(t\\
&\quad=1)|A\\
&=a,V,W^{\mathrm{c}}(t\\
&=0),W^{\mathrm{c}}(t\\
&=1))-E(Y^{\mathrm{c}}(t\\
&\quad=0)|A\\
&=a,V,W^{\mathrm{c}}(t\\
&=0),W^{\mathrm{c}}(t\\
&\quad=1))\\
&=E(Y^{\mathrm{c}}(t\\
&\quad=1)|A\\
&=a,V,W^{\mathrm{c}}(t\\
&\quad=1))\\
&-E(Y^{\mathrm{c}}(t\\
&\quad=0)|A\\
&\quad=a,V,W^{\mathrm{c}}(t=0))
\end{aligned}
$$

As with estimand II, the first equality in the identifiability result holds under the RA(9). The last equality holds under assumption (11) (i.e. by substituting t=1 and t = 0 for t), giving us a third estimand for the ATE:

$$\Psi^{\mathrm{III}}(P_0)=E_{V,W(t)}\begin{pmatrix}E(Y^{\mathrm{c}}(t=1)|A=1,V,W^{\mathrm{c}}(t=1))-E(Y^{\mathrm{c}}(t=0)|A=1,V,W^{\mathrm{c}}(t=0))\\-(E(Y^{\mathrm{c}}(t=1)|A=0,V,W^{\mathrm{c}}(t=1))-E(Y^{\mathrm{c}}(t=0)|A=0,V,W^{\mathrm{c}}(t=0)))\end{pmatrix} \quad (12)$$

We refer to this final estimand as the pooled outcome estimand. However, additional restrictions on the allowed data distribution are needed for this identifiability result to hold. Starting with the SCM established for the change score estimand (Ψ^{II}), we work through the model separately at each time point. At time t=1, assumption (11) becomes:

$$E_{V,W(t=1),W(t=0)}(Y^{\mathrm{c}}(t=1)|A=a,V,W^{\mathrm{c}}(t=0),W^{\mathrm{c}}(t=1))=E_{V,W(t=1)}(Y^{\mathrm{c}}(t=1)|A=a,V,W^{\mathrm{c}}(t=1)),$$

which will hold if $Y^{\mathrm{c}}(t = 1)$ is independent of $W^{\mathrm{c}}(t =0)$ given V, A, and $W^{\mathrm{c}}(t=1)$. We can use the DAG shown in Figure 7 to check whether our SCM implies this conditional independence. Under our current model, assumption (11) fails at t = 1 because of two

unblocked paths: the direct path from $W^c(t = 0)$ to $Y^c(t = 1)$ (label (vii) in Figure 7); and the paths through collider A (i.e. $W^c(t= 0) - C \rightarrow Y^c(t=1)$ label (viii) in Figure 7). Therefore, for assumption (11) to hold at $t = 1$, we need to add two new exclusion restrictions: that $W^c(t = 0)$ does not affect $Y^c(t=1)$ and does not affect A (see Figure 8).

Similarly, at time $t = 0$, assumption (11) becomes:

$$E_{V,W(t=1),W(t=0)}(Y^c(t=0)|A=a,V,W^c(t=0),W^c(t=1))=E_{V,W(t=0)}(Y^c(t=0)|A=a,V,W^c(t=0))$$

and we verify with a DAG that our SCM implies $Y^c(t= 0)$ is independent of $W^c(t= 1)$ given V, A, and $W^c(t = 0)$ (Figure 9). No additional exclusion restrictions are required.

Note that we cannot add any arrows back that were removed for estimand II (i.e. $Y^c(t =0)$ cannot affect A, $W^c(t = 1)$ or $Y^c(t = 1)$). Under the additional restriction assumptions that $W^c(t =0)$ does not affect A and $Y^c(t = 1)$, our causal target parameter, the ATE, is equivalent to estimand III. In settings where background knowledge makes it plausible to assume this more restrictive causal model, the pairing of $W^c(t)$ and $Y^c(t)$ at time t may result in an efficiency gain.

Illustration of identifiability results using simulated data

In this section, we present a series of simulations to illustrate the reliance of estimand Ψ^I and the difference-in-differences estimands Ψ^{II} and Ψ^{III} on distinct identifiability assumptions (the code is available in the Appendix A). As with the DAGs, we excluded the observed village factors, V, from the simulations. We present eight scenarios based on different SCMs represented by the DAGs in the previous section. In all cases, $Y^c(t)$, $W^c(t)$ and C are continuous, normally distributed and a function of additive linear terms. Treatment variable, A, is dichotomous and the true causal parameter of interest, the ATE, has a value of 1. For each scenario and estimand, we estimated the true value of the estimand using simulation based on a sample of 100,000 observations (using fits of correctly specified linear regressions for each conditional expectation). These estimates are reported in Table 2.

The first simulation is based on the starting SCM for the post-treatment estimand (Ψ^I) represented in Figure 1. Under this model, RA(1) holds, and the true value of the estimand is equivalent whether the outcome is defined as $Y^c(t= 1)$ or Y^θ (Figure 2 and RA(4)) and is equal to the target parameter value of 1 (simulation #1, Table 2). However, when we introduce an unmeasured confounder, C, in the second simulation, RA(1) and RA(4) no longer hold and the estimand diverges from the target causal parameter (simulation #2). This result is in keeping with a backdoor pathway being open from A to outcome $Y^c(t = 1)$ through C (path (i) in Figure 3) or with dependence between Y^θ and A through $U_{Y(t=0)}$ and confounder C (path (ii) in Figure 4).

The change score estimand (Ψ^{II}) diverges from the true value of the ATE when not conditioning on $Y^c(t=0)$ (simulation #3) because this opens up new pathways from A to Y^θ (paths (iii) to (vi) in Figure 5). After adding the additional exclusion restrictions for

estimand II in the fourth simulation (i.e. Figure 6), the true value of the estimand again equals the target casual parameter value (simulation #4). However, in simulation #5, we add that $Y^c(t=0)$ affects A into the previous scenario for estimand II. In this fifth scenario, estimand II diverges from the ATE.

The sixth simulation represents the model for our pooled outcome estimand (Ψ^{III}), where at time $t=1$, $Y^c(t=1)$ is not independent of $W^c(t=0)$ given V, A, and $W^c(t=1)$ (Figure 7). As expected, estimand III diverges from the ATE (simulation #6). However, when the paths from $W^c(t=0)$ to A and $Y^c(t=1)$ are removed (Figure 8), estimand III is equal to the ATE (simulation #7). Finally, in simulation #8, we add that $Y^c(t=0)$ affects A into the previous scenario for estimand III, and the estimate once again diverges from the truth. As with simulation #5, this last simulation demonstrates that even if we can accept all the other exclusion restrictions for estimand III, we still must be willing to accept that $Y^c(t=0)$ does not affect A for the difference-in-differences estimands to equal the target parameter.

In summary, the above simulations show that when there is an unmeasured confounder, the post-treatment estimand is generally not equal to the ATE whereas the change score and pooled outcome estimands might be, but only under additional assumptions. We further illustrate that even with an additive constant confounder C, the latter two estimands (Ψ^{II} and Ψ^{III}) may still diverge substantially from the ATE if $Y^c(t=0)$ affects A (i.e. $Y^c(t=0)$ is a confounder), as well as if additional assumptions fail to hold.

Estimation methods

To estimate the ATE of the nutrition program on children's mean weight-for-age in a community, we used targeted maximum likelihood estimation (TMLE) for each of the three estimands [4, 25]. TMLE is a doubly robust estimator with important advantages over more commonly used estimators, such as parametric regression or inverse probability of treatment weights (IPTW or propensity score weighting) [26, 27]. Because the Madagascar evaluation was an *ex-post facto* quasi-experimental design, we wanted an estimator that would do the best job possible of adjusting for confounding from covariate imbalance across treatment groups (by chance or by design of the program roll-out).

TMLE involves estimation of both the conditional mean of the outcome given treatment and covariates, $\overline{Q_0}$, and the conditional probability of treatment given covariates, g_0, in estimating a causal effect. The initial estimator of $\overline{Q_0}$ is updated in a fluctuation procedure using a "clever covariate," which is a function of the treatment mechanism, g_0 [4]. The updated estimates of the predicted values of the outcome under each treatment condition are then used to obtain an estimate of the statistical target parameter of interest. In this way, TMLE removes all asymptotic residual bias of the initial estimator for the target parameter, as long as we have a consistent estimator for $\overline{Q_0}$ or g_0 [4].

$\overline{Q_0}$ for estimands I, II and III is defined respectively, as:

$$\text{For}\Psi^{\text{I}}{:}\overline{Q_0}{=}E_0[Y^{\text{c}}(t{=}1)|A{=}\text{a},V,W^{\text{c}}(t{=}0),Y^{\text{c}}(t{=}0),W^{\text{c}}(t{=}1)],$$
$$\text{For}\Psi^{\text{II}}{:}\overline{Q_0}{=}E_0[Y^{\theta}|A{=}\text{a},V,W^{\text{c}}(t{=}0),W^{\text{c}}(t{=}1)],\text{and}$$
$$\text{For}\Psi^{\text{III}}{:}\overline{Q_0}{=}E_0[Y^{\theta}(t)|A{=}\text{a},V,W^{\text{c}}(t),t]$$

and the corresponding g_0 is defined as:

$$\text{For}\Psi^{\text{I}}{:}g_0(a|V,W^{\text{c}}(t{=}0),Y^{\text{c}}(t{=}0),W^{\text{c}}(t{=}1)){=}P_0(A{=}\text{a}|V,W^{\text{c}}(t{=}0),Y^{\text{c}}(t{=}0),W^{\text{c}}(t{=}1))$$
$$\text{For}\Psi^{\text{II}}{:}g_0(a|V,W^{\text{c}}(t{=}0),W^{\text{c}}(t{=}1)){=}P_0(A{=}a|V,W^{\text{c}}(t{=}0),W^{\text{c}}(t{=}1))$$
$$\text{For}\Psi^{\text{III}}{:}g_0(a|V,W^{\text{c}}(t),t){=}P_0(A{=}\text{a}|V,W^{\text{c}}(t),t)$$

TMLE was implemented using R package “Targeted Maximum Likelihood Estimation” version 1.2.0-4 [28]. In order to avoid unsubstantiated parametric assumptions on the data-generating process, we fit both g_0 and $\overline{Q_0}$ using SuperLearner (“SuperLearner Prediction” version 2.0-10, see Appendix B for a description), a machine-learning algorithm based on 10-fold cross-validation [29]. Candidate SuperLearner (SL) algorithms that were included for fitting *g* were generalized linear models, Bayesian linear models, generalized additive models, step-wise regression, k-nearest neighbors and neural networks. Candidate algorithms that were included for fitting Q were generalized linear models, Bayesian linear models, generalized additive models, step-wise regression and polynomial spline regression. Standard errors were estimated using a non-parametric bootstrap with 200 replications. For each bootstrap sample, 410 communities were sampled with replacement and estimates for each estimand obtained before drawing the next sample. The confidence intervals (CI) were calculated assuming a normal distribution for the estimator, as well as by ordering the bootstrap estimates and taking the 2.5th and 97.5th percentile values. We also present robust influence curve based confidence intervals using the TMLE package [28]. Finally, we compared TMLE to two other methods of estimation: linear main term regression and inverse probability of treatment weighting [26, 27].

Estimation results using observed data

In a simulation under which all identifying assumptions held, the TMLE with SL estimator for each estimand was unbiased and the IC based 95% CI achieved nominal coverage (see Appendix C). Estimates for the point treatment effect from the observed data and their corresponding confidence intervals are shown in Table 3. The point estimates represent a difference in community mean weight-for-age z-score. A unit change of one is equivalent to one standard deviation (SD) above the mean weight for the reference standard (i.e. a population of well-nourished and healthy children of the same age and gender).

For estimand I, the estimate of the ATE obtained with TMLE is less than a tenth of a standard deviation, but statistically significant at the 5% level for two of the three CI's (β=0.066, influence curve-based CI: 0.009, 0.123). The differences in CI are small. The point estimate for estimand II is much larger than estimand I and statistically significant (β = 0.255, influence curve-based CI: 0.165, 0.346). As with estimand II, the point estimate for

estimand III is relatively large and statistically significant ($\beta = 0.274$, influence curve-based CI: 0.184, 0.365).

Differences between estimates for any given estimand using TMLE with SuperLearner, inverse probability of treatment weighting and linear main term regression were very small compared to the differences between estimands (data not shown).

Discussion

Pre-post program evaluations (with data from treatment and control groups) present investigators with multiple approaches for identifying the causal effect of the program, each of which relies on a different set of assumptions. In this paper, we consider an existing program evaluation with pre-post data in order to illustrate trade-offs implied by alternative approaches to identifiability. We further contrast the results obtained by applying an efficient double robust estimator (TMLE) to observed data from the evaluation in order to estimate each of the corresponding estimands.

We show with simulated data that when the outcome is defined as the post-treatment value, $Y^c(t = 1)$, under the key assumption of no unmeasured confounding, the simple post-treatment estimand (Ψ^I) equals the ATE (our target casual parameter). If an unmeasured factor, C, is introduced that confounds the relationship between treatment and outcome, the Ψ^I and the ATE diverge. Since unmeasured confounding is a realistic scenario in observational studies, it is not surprising that a difference-in-differences approach is often favored to try to address this issue. A differencing model is advantageous in that it "subtracts out" the effect of unmeasured confounders with a constant additive effect on the outcome at the two time points. The commonly accepted identifying assumption for the difference-in-differences estimand is a RA known as the parallel trend assumption. However, additional assumptions are necessary for the difference-in-differences estimand to equal the ATE. We discuss a sufficient set of such assumptions in a non-parametric structural equation model, namely that the lagged outcome, $Y^c(t = 0)$, does not affect treatment, A, the post-treatment covariates, $W^c(t= 1)$, or the post-treatment outcome, $Y^c(t= 1)$. These are very strong assumptions about the lagged outcome. Under conditions where these restrictions do not hold, difference-in-differences estimands (Ψ^{II} and Ψ^{III}) have the potential to diverge further from the wished for causal effect than the post-treatment estimand adjusting for all baseline covariates, even in the presence of an unmeasured confoun-der with a constant additive effect. The exclusion restrictions become more numerous for the model that pools the outcome from both time periods (Ψ^{III}).

When estimating the ATE using the observed data and TMLE, we obtained a small, although statistically significant, point estimate for the post-treatment estimand, Ψ^I: less than one tenth of a standard deviation in mean weight-for-age z-score. In contrast, the point estimates of the ATE for the differencing estimands are much larger at 0.26–0.27 SD in mean weight-for-age z-score. Drawing conclusions about the impact of the nutrition program requires us to evaluate the plausibility of the identifying assumptions underlying each approach. Do the larger effect estimates of the differencing estimands represent an estimate of the ATE that is less biased by some unmeasured, time-independent, confounder

(e.g. community dispersion) or an estimate that is more biased due to confounding by baseline mean weight-for-age z-score ($Y^c(t=0)$)?

In 1997, the Malagasy government based selection into the treatment group in part on district-level prevalence of moderate underweight in children less than 5 years, aggregated up from community data. This background knowledge suggests that we should condition on $Y^c(t=0)$. However, the *ex-post facto* evaluation of program impact was designed and implemented in 2004. Differencing models are often applied to data from serial cross-sections of different persons from the same communities separated in time by many years [24]. It is possible under certain conditions that the pre-treatment outcome and covariates do not directly affect the post-treatment outcome and covariates and are associated with post-treatment outcome and covariates due only to fixed community-level factors that affect both. In other words, $Y^c(t=0)$ may be predictive of $Y^c(t=1)$, but only due to shared common causes C, $W^c(t=0)$, V, or $W_c(t=1)$, given that the cross-sectional surveys were administered 7 years apart in Madagascar. Therefore, if we accept that Ψ^{II} or Ψ^{III} is equal to our parameter of interest, then we accept that any residual variation in $Y^c(t=0)$ not explained by $W^c(t=0)$ and V has only a minimal influence on Y^{θ}.

In the econometrics literature, other authors have shown using parametric models that under certain assumptions and conditions for selection into treatment, a difference-in-differences model and a model conditional on the lagged outcome will provide upper and lower bound estimates of the causal effect of interest [23, 24]. We briefly consider whether our two estimands, the difference-in-differences estimand, Ψ^{III}, or the post-treatment estimand (referred to as a lagged-outcome estimand in this literature), Ψ^{I}, can be interpreted similarly under our non-parametric causal model, given our contextual knowledge of factors that influenced selection into treatment in this study.

In the Madagascar study, both contextual knowledge and observed associations strongly suggest that the treatment was differentially assigned to villages with a lower lagged outcome (lower mean weight-for-age in a community or higher prevalence of underweight). In this context, failure to control for the lagged outcome is expected to result in a differences-in-differences estimand that overestimates the effect of interest. Interpretation of Ψ^{III} as a true upper bound, however, would require that all other assumptions hold for the difference-in-differences estimand to equal the causal effect of interest, or that any which fail are similarly contributing to an overestimate of the effect.

On the other hand, as noted by Guryan in the context of his econometric model [24], a lagged-outcome estimator may under- or over-estimate the true effect depending on the direction of unmeasured confounding. In the Madagascar study, unmeasured confounding of the post-treatment estimand is unlikely to have resulted in a substantial overestimate of the true effect, as the post-treatment estimate is already near zero and there is not a good basis for suspecting a meaningful detrimental effect of the intervention. In other words, simply based on a priori knowledge, an effect estimate of zero is a plausible lower bound. It remains possible, of course, that the post-treatment estimand resulted in an underestimate of the true effect due to unmeasured confounding. For example distance of the village from a qualified non-profit organization to monitor the program may have both decreased

probability of assignment to treatment and independently improved outcomes; while some of this confounding would be expected to be captured though adjustment for the lagged outcome, residual negative confounding might remain.

Conclusions

In summary, our results highlight important trade-offs between approaches to identifiability in the context of evaluating an intervention with pre-post data. We are confronted with a bias trade-off between a single post-treatment estimand that conditions on the pre-treatment outcome (a measured confounder) but assumes no unmeasured confounders, and two difference-in-differences estimands that address certain types of unmeasured confounders but do not condition on the pre-treatment outcome. Note that none of the estimands account for time-varying and non-linear unmeasured confounders. The equivalence of the estimands to the casual effect of interest relies on assumptions that cannot be empirically verified; we require expert knowledge about the process that generated the observed data before we can choose one over the other.

If our knowledge is sufficient to accurately represent the underlying data-generating process using a causal model, then our casual model may help us choose between estimands (e.g. to decide whether the post-treatment estimand is closer to the ATE than the differencing estimands). In the case of the Madagascar intervention, our knowledge is unfortunately not sufficient to say definitively which, if any, of the SCM that we considered accurately describe the true data-generating process, nor which corresponding estimand is closer to our causal target parameter. If the required assumptions hold for the models described by Guryan and others, the truth lies somewhere between zero effect and a moderate positive effect of the program, which was sufficient to warrant further monitoring of the program.

Finally, if we have strong evidence that (a) there is important unmeasured confounding and that (b) the data do not support any other assumptions on which our identifiability results rely, then the target parameter is not identifiable. We cannot disregard this evidence; we risk obtaining a biased estimate, which has important implications for future funding of a program. The threat to validity from selecting a causal model without understanding the underlying assumptions transcends our work and is applicable to any evaluation of an intervention.

Acknowledgments

This work was made possible by Dr. Emanuela Galasso, Senior Economist at the World Bank, who gave the authors access to the Madagascar study data and information. Additional guidance and support were provided by Drs. Ira B. Tager, Lia H. Fernald, and Ray Catalano at the University of California, Berkeley.

Appendix A: R Code for simulations for Table 2

```
#————————————————————

# In all, we assume W1 is not affected by A, and exclude observed exogenous variables, V

#————————————————————
```

```
set.seed(100)

n <- 100000

C <-rnorm(n,0,1)

W0<-rnorm(n,0,1)

#————————————————————

# Run 1: Example for figure 3.1: estimand I, controlling for Y0

# No unmeasured confounding C

Y0 <-rnorm(n,0.5*W0,1)

A <-rbinom(n,1,1/(1 + exp(-0.5*W0-0.5*Y0)))

W1 <-rnorm(n,W0+Y0,1)

Y1 <-rnorm(n,W0+2*Y0+A+W1,1)

est1 <- glm(Y1∼A+W0+W1+Y0)

#————————————————————————————————————

# Run 2: Example for figure 3.3: estimand I

# Introduce unmeasured confounder C that affects Y(0), Y(1) and A

Y0<-rnorm(n,0.5*W0+C,1)

A <-rbinom(n,1,1/(1 + exp(-0.5*W0-0.5*Y0-0.5*C)))

W1 <-rnorm(n,W0+Y0,1)

Y1 <-rnorm(n,W0+Y0+A+W1+C,1)

est2 <- glm(Y1∼A+W0+W1+Y0)

#————————————————————————————————————

# Run 3: Example for figure 3.5: estimand II, not controlling for Y(0)

# Unmeasured confounder C

Y0<-rnorm(n,0.5*W0+C,1)

A <-rbinom(n,1,1/(1 + exp(-0.5*W0-0.5*Y0-0.5*C)))

W1 <-rnorm(n,W0+Y0,1)
```

```
Y1 <-rnorm(n,W0+Y0+A+W1+C,1)

Yd <-Y1-Y0

est3 <- glm(Yd~A+W0+W1)

#————————————————————————————————————

# Run 4: Example for figure 3.6: estimand II, not controlling for Y(0)

# Confounder C, assume Y(0) does not affect A, W(1), or Y(1); i.e., no confounding by Y(0)

Y0<-rnorm(n,0.5*W0+C,1)

A <-rbinom(n,1,1/(1 + exp(-0.5*W0-0.5*C)))

W1 <-rnorm(n,W0,1)

Y1 <-rnorm(n,W0+A+W1+C,1)

Yd <-Y1-Y0

est4 <- glm(Yd~A+W0+W1)

#————————————————————————————————————

# Run 5: Example adding Y(0) affects A into run 4

Y0<-rnorm(n,0.5*W0+C,1)

A <-rbinom(n,1,1/(1 + exp(-0.5*W0-0.5*Y0-0.5*C)))

W1 <-rnorm(n,W0,1)

Y1 <-rnorm(n,W0+A+W1+C,1)

Yd<-Y1-Y0

est5 <- glm(Yd~A+W0+W1)

#————————————————————————————————————

# Run 6: Example for figure 3.7: estimand III

# Confounder C, assume Y(0) does not affect A, W(1), or Y(1); i.e., no confounding by Y(0)

# Assumption (11) but W(0) affects A and Y(1)

Y0 <-rnorm(n,0.5*W0+C,1)

A <-rbinom(n,1,1/(1 +exp(-0.5*W0-0.5*C)))
```

```
W1<-rnorm(n,W0,1)

Y1<-rnorm(n,W0+A+W1+C,1)

# Reshape wide to long

id <- paste("id", 1:n, sep = "")

data_wide <- data.frame(id,C,A,W0,Y0,W1,Y1)

data_long <- reshape(data_wide,
      varying=4:7,
      idvar = "id",
      direction = "long",
      timevar="T",
      new.row.names=NULL,
      sep = "")
est6 <- glm(Y∼A+W+T+A*T,data=data_long)

#————————————————————————————————————

# Run 7: Example for figure 3.8: estimand III

# Confounder C, assume Y(0) does not affect A, W(1), or Y(1); i.e., no confounding by Y(0)

# Assumption (11) and W(0) does not affect A or Y(1)

Y0 <-rnorm(n,0.5*W0+C,1)

A <-rbinom(n,1,1/(1 +exp(-0.5*C)))

W1<-rnorm(n,W0,1) Y1<-rnorm(n,A+W1+C,1)

# Reshape wide to long

id <- paste("id", 1:n, sep = "")

data_wide <- data.frame(id,C,A,W0,Y0,W1,Y1)

data_long <- reshape(data_wide,
      varying=4:7,
      idvar = "id",
      direction = "long",
      timevar="T",
```

```
    new.row.names=NULL,
    sep = "")
est7 <- glm(Y∼A+W+T+A*T,data = data_long)
#————————————————————————————————————
# Run 8: Example adding Y(0) affects A into run 7
Y0 <-rnorm(n,0.5*W0+C,1)
A <-rbinom(n,1,1/(1 +exp(-0.5*Y0-0.5*C)))
W1<-rnorm(n,W0,1)
Y1<-rnorm(n,A+W1+C,1)
# Reshape wide to long
id <- paste("id", 1:n, sep= "")
data_wide <- data.frame(id,C,A,W0,Y0,W1,Y1)
data_long <- reshape(data_wide,
    varying = 4:7,
    idvar="id",
    direction = "long",
    timevar="T",
    new.row.names = NULL,
    sep ="")
est8 < - glm(Y∼A + W + T + A*T,data = data_long)
#————————
est_all <-rbind(est1$coeff["A"],est2$coeff["A"],est3$coeff["A"],est4$coeff["A"],
est5$coeff["A"], est6$coeff["A:T"],est7$coeff["A:T"],est8$coeff["A:T"])
est_all
```

Appendix B: SuperLearner

SuperLearner (SL) [30] is a non-parametric, machine-learning tool that "learns" from the observed data by using a candidate set of algorithms (or estimators) and a pre-specified loss function that assigns a measure of performance to each of the algorithms. Briefly, there are three key components to SL:

1. SL uses a library of algorithms for prediction. The algorithms can be diverse, simple (i.e. logistic regression), complex (i.e. neural nets), numerous and can include user-defined algorithms.
2. The predictive performance of each algorithm is assessed using V-fold cross-validation. Cross-validation involves partitioning the sample into a user-specified number of training and validation sets. A training set is used to construct the candidate estimators (i.e. fit the regression) and the corresponding validation set is then used to assess the performance (i.e. estimate the risk) of the candidate algorithms. The validation set rotates by the number of partitions such that each set is used as the validation set once. Risk is defined using a loss function, for example, if we use the squared error loss function then our estimate of the risk corresponds to the estimated mean squared error loss on the validation sets. The "best" algorithms typically have the smallest empirical risk averaged over all the validation sets.
3. The library of algorithms is augmented with new algorithms, which are weighted averages of the algorithms from the previous step. The weighted algorithm with the smallest cross-validated risk is the "super learner" estimator and is expected to outperform any single algorithm. (Note that we can include a parametric model in the SuperLearner library.)

Appendix C: simulation with the TMLE and SL estimator

We checked the performance of the TMLE and SL estimator using simulated data (generated as in Appendix A) and a sample size of 410 (the same size as the observed data), where the true value of the estimand was 1 and all identifying assumptions held for each estimand. The simulation was repeated 5,000 times to show that the TMLE_SL estimator was unbiased and the IC based 95% CI achieved nominal coverage (results given in below table).

	Mean	Variance	Bias	MSE	Coverage†
Estimand I	1.00	0.013	0.0001	0.013	0.94
Estimand II	1.00	0.021	0.0001	0.021	0.94
Estimand III	1.01	0.020	0.0051	0.020	0.95

Note:

†Proportion of runs where the 95% influence curve-based CI contains the true value.

References

1. Engle PL, Fernald LC, Alderman H, Behrman J, O'Gara C, Yousafzai A, et al. Strategies for reducing inequalities and improving developmental outcomes for young children in low-income and middle-income countries. Lancet. 2011; 378:1339–53. [PubMed: 21944378]
2. Heckman, JJ.; Vytlacil, EJ. Chapter 70 econometric evaluation of social programs, part I: causal models, structural models and econometric policy evaluation. In: Heckman, J.; Leamer, E., editors. Handbook of econometrics. Amsterdam: Elsevier; 2007. p. 4779-874.

3. Petersen ML, van der Laan MJ. Causal models and learning from data: integrating causal modeling and statistical estimation. Epidemiology. 2014; 25:418–26. [PubMed: 24713881]
4. van der Laan, MJ.; Rose, S. Targeted learning: causal inference for observational and experimental data. Berlin, Heidelberg, New York: Springer; 2011.
5. Pearl J. An introduction to causal inference. Int J Biostat. 2010; 6(2) Article 7.
6. Imai, K. Causal inference lecture notes: causal inference with repeated measures in observational studies. Department of Politics; Princeton University: 2008.
7. Maris E. Covariance adjustment versus gain scores – revisited. Psychol Methods. 1998; 3:309–27.10.1037/1082-989x.3.3.309
8. Gertler, PJ.; Martinez, S.; Premand, P.; Rawlings, LB.; Vermeersch, CM. Impact evaluation in practice. Washington, DC: The International Bank for Reconstruction and Development/The World Bank; 2011.
9. Lord FM. The measurement of growth. Educ Psychol Meas. 1956; 16:421–37.10.1177/001316445601600401
10. Meyer BD. Natural and quasi-experiments in economics. National bureau of economic research technical working paper series no 170. 1994
11. Pearl J. Causal diagrams for empirical research. Biometrika. 1995; 82:669–88.10.1093/biomet/82.4.669
12. UNICEF. [cited February 9 2011] Info by country: Madagascar statistics 2011. 2011. Available at: http://www.unicef.org/infobycountry/madagascar_statistics.html
13. Black RE, Allen LH, Bhutta ZA, Caulfield LE, de Onis M, Ezzati M, et al. Maternal and child undernutrition: global and regional exposures and health consequences. Lancet. 2008; 371:243–60. [PubMed: 18207566]
14. Galasso, E.; Umapathi, N. World Bank Policy Research Working Paper Series No 4424. Washington, DC: 2007. Improving nutritional status through behavioral change: lessons from Madagascar.
15. Robins JM, Hernán MÁ, Brumback B. Marginal structural models and causal inference in epidemiology. Epidemiology. 2000; 11:550–60. [PubMed: 10955408]
16. Rubin DB. Statistics and causal inference: comment: which ifs have causal answers. J Am Stat Assoc. 1986; 81:961–2.
17. Pearl J. On the consistency rule in causal inference: an axiom, definition, assumption, or a theorem? Epidemiology. 2010; 21(6):872–875. [PubMed: 20864888]
18. Greenland S, Pearl J, Robins JM. Causal diagrams for epidemiologic research. Epidemiology. 1999; 10:37–48. [PubMed: 9888278]
19. Petersen ML, Porter K, Gruber S, Wang Y, van der Laan MJ. Diagnosing and responding to violations in the positivity assumption. 2010
20. Wang, Y.; Petersen, ML.; Bangsberg, D.; van der Laan, MJ. Diagnosing bias in the inverse probability of treatment weighted estimator resulting from violation of experimental treatment assignment. UC Berkeley Division of Biostatistics Working Paper Series. 2006. Available at: http://www.bepress.com/ucbbiostat/paper211
21. Rubin DB, Stuart EA, Zanutto EL. A potential outcomes view of value-added assessment in education. J Educ Behav Stat. 2004; 29:103–16.
22. Allison, PD. Change scores as dependent variables in regression analysis. In: Clogg, Clifford, editor. Sociological methodology. Oxford, Cambridge, MA: Basil Blackwell; 1990. p. 93-114.
23. Angrist, J.; Pischke, JS. Mostly harmless econometrics: an empiricist's companion. Princeton, New Jersey: Princeton University Press; 2009.
24. Guryan J. Desegregation and black dropout rates. Am Econ Rev. 2004; 94:919–43.
25. van der Laan MJ, Rubin D. Targeted maximum likelihood learning. Int J Biostat. 2006; 2(1) Article 11.
26. Hernán MA, Robins JM. Estimating causal effects from epidemiological data. J Epidemiol Commun Health. 2006; 60:578–86.
27. Hirano K, Imbens GW, Ridder G. Efficient estimation of average treatment effects using the estimated propensity score. Econometrica. 2003; 71:1161–89.

28. Gruber S, van der Laan M. Tmle: an R package for targeted maximum likelihood estimation. J Stat Softw. 2012; 51(13) Available at: http://www.jstatsoft.org/v51/i13/paper.

29. Polley, E.; van der Laan, M. SuperLearner: super learner prediction. Available at: http://cran.r-project.org/web/packages/SuperLearner/index.html

30. Polley, EC.; Rose, S.; van der Laan, MJ. Super learning. In: van der Laan, MJ.; Rose, S., editors. Targeted learning: causal inference for observational and experimental data. Berlin, Heidelberg, New York: Springer; 2011.

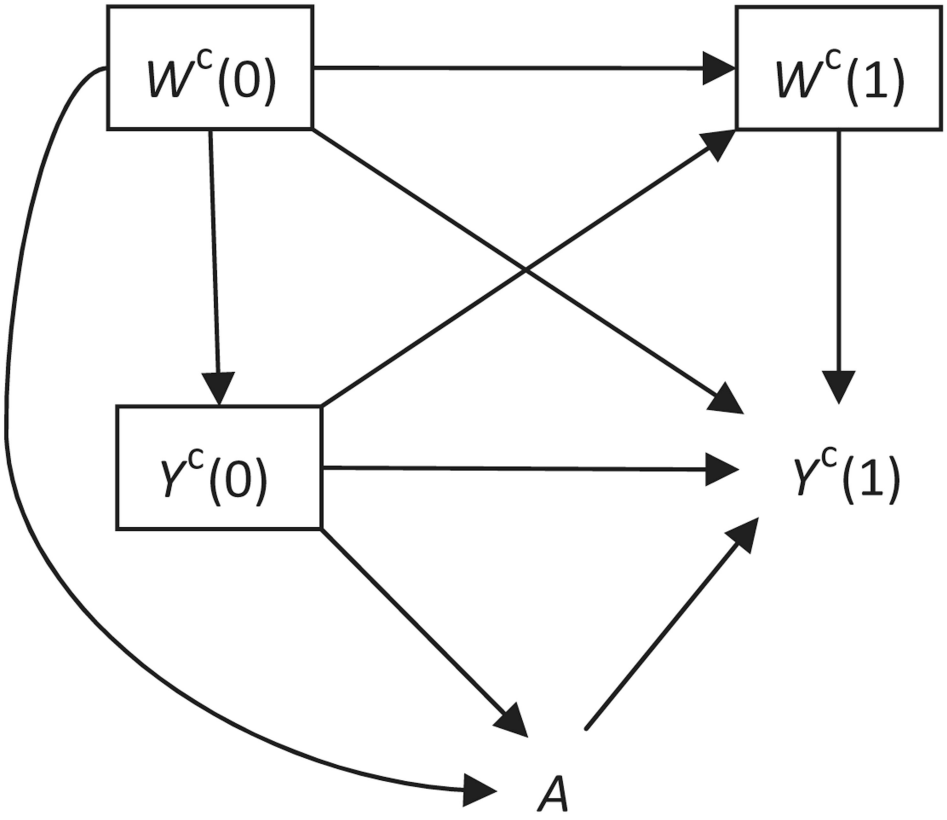

Figure 1.
DAG illustrating that post-treatment outcome, $Y^c(1)$, is independent of treatment, A, given lagged outcome, $Y^c(0)$, pre- and post-treatment covariates, $W^c(0)$ and $W^c(1)$, and exogenous covariates, V (not shown). There are no unmeasured confounders.

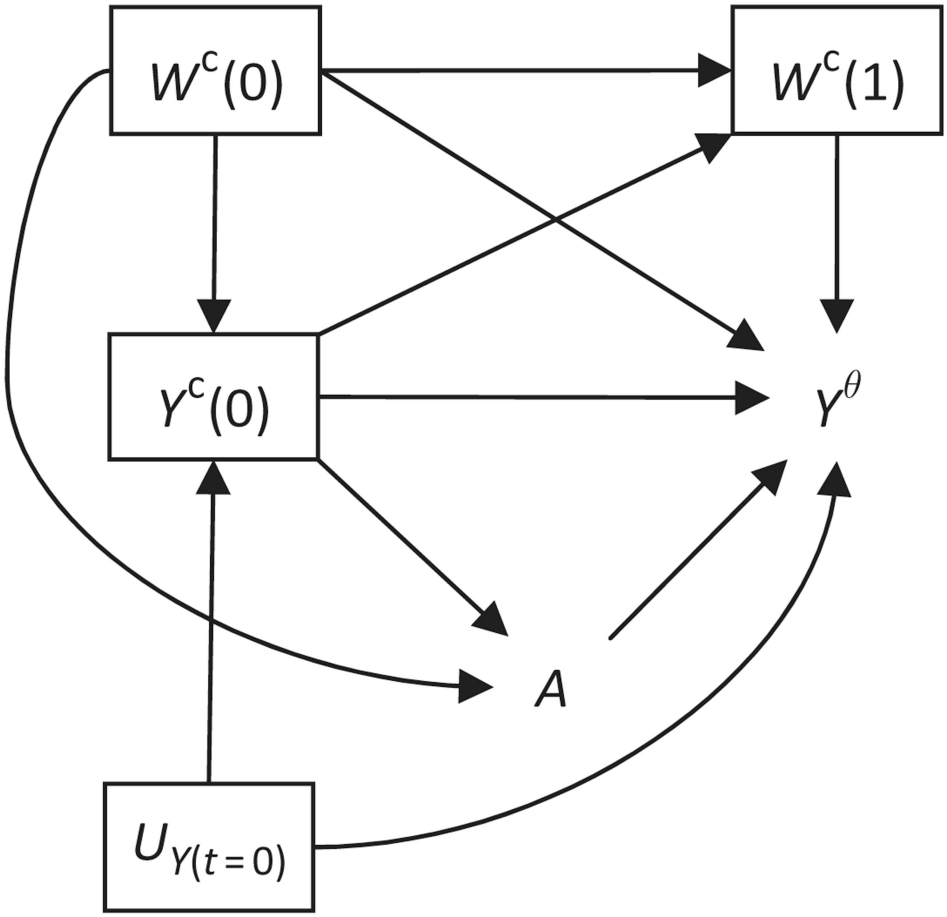

Figure 2.

DAG illustrating that pre-post change outcome, Y^θ, is independent of treatment, A, given the lagged outcome, $Y^c(0)$, pre- and post-treatment covariates, $W^c(0)$ and $W^c(1)$, and exogenous covariates, V (not shown). There are no unmeasured confounders.

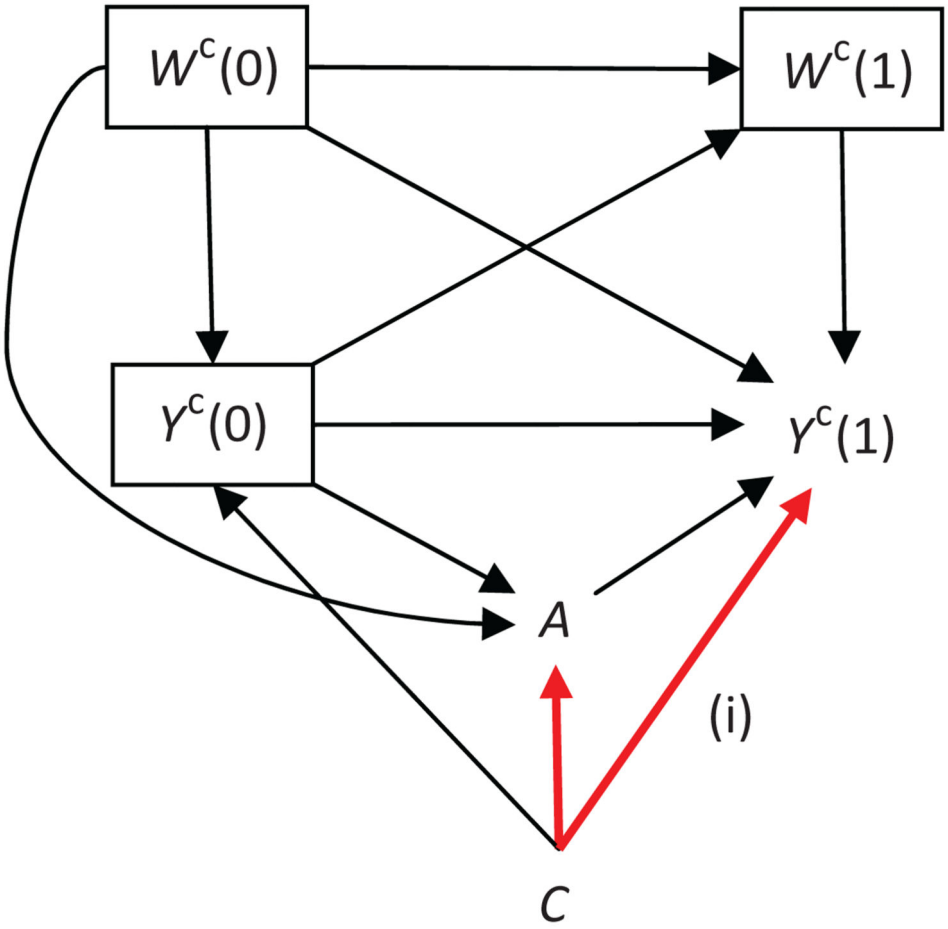

Figure 3.
DAG illustrating that an unblocked path (i) is opened from treatment, *A*, to the post-treatment outcome, $Y^c(1)$, in the presence of unmeasured confounder, *C*.

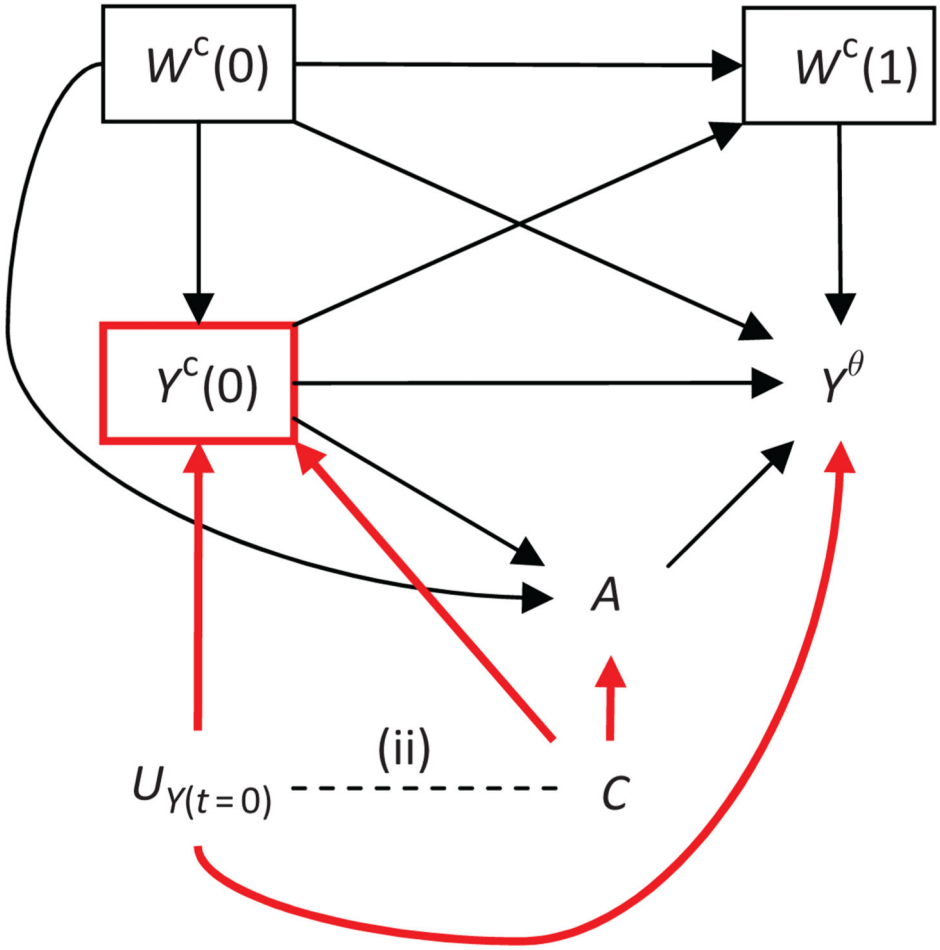

Figure 4.

DAG illustrating by conditioning on pre-treatment outcome, $Y^c(0)$, in the presence of unmeasured confounder, C, that an unblocked path (ii) is opened from treatment, A, to the pre-post change outcome, Y^θ, through C and exogenous variable $U_{Y(t=0)}$.

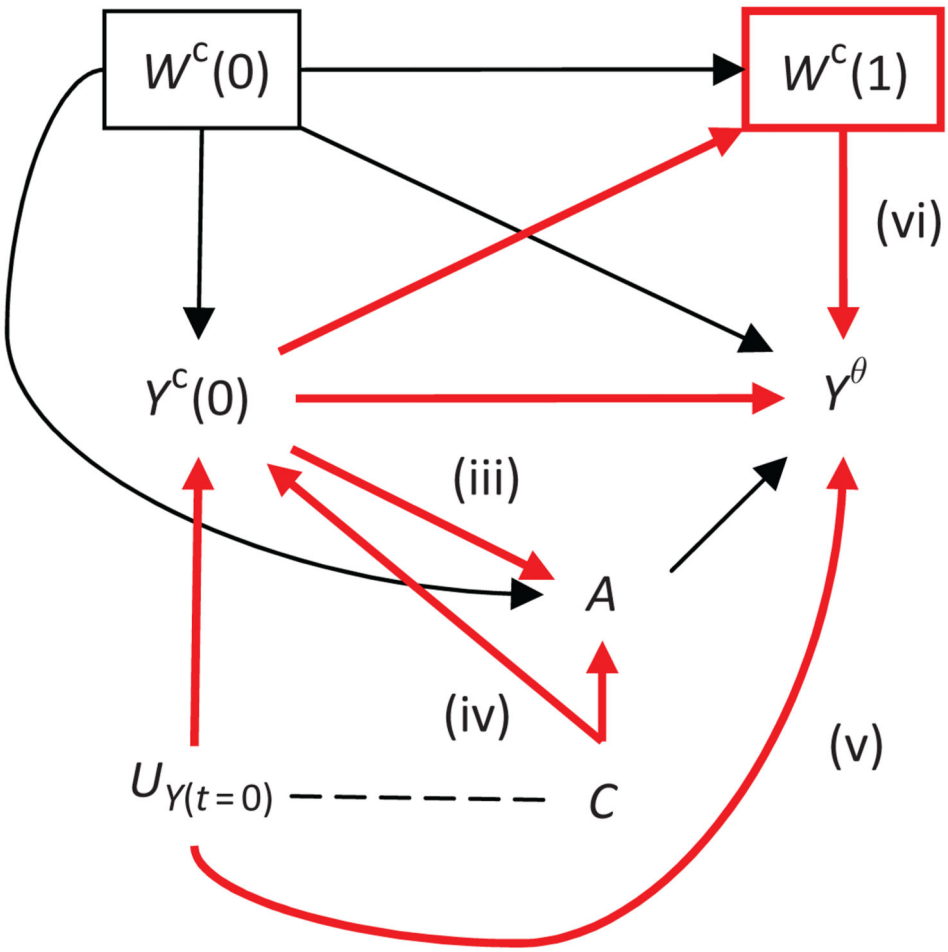

Figure 5.

DAG illustrating by not conditioning on lagged outcome, $Y^c(0)$, in the presence of unmeasured confounder, C, that multiple unblocked paths are opened from treatment, A, to the pre-post change outcome, Y^θ, through: (iii) lagged outcome, $Y^c(0)$; (iv) confounder, C; (v) exogenous $U_{Y(t=0)}$; and (vi) collider covariates $W^c(1)$.

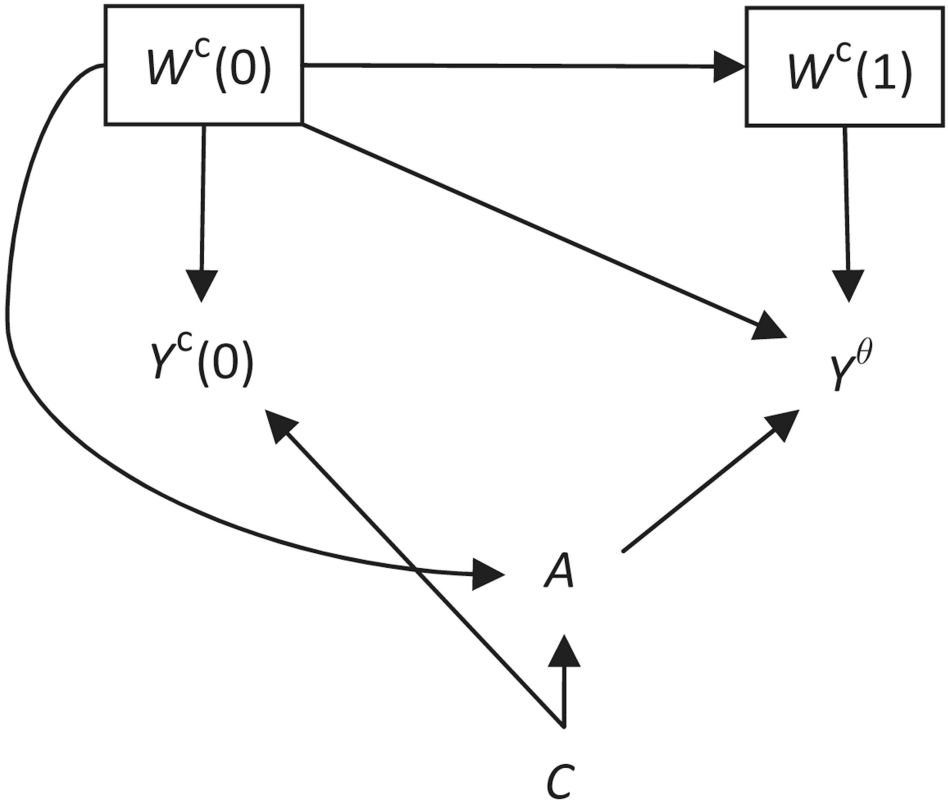

Figure 6.

DAG illustrating exclusion restrictions on lagged outcome Y^c (0) for the pre-post change outcome, Y^θ, in the presence of unmeasured confounder, C. Y^θ, is independent of treatment, A, given covariates, only if $Y^c(0)$, does not affect treatment, A, post-treatment outcome, $Y^c(1)$, and post-treatment covariates, $W^c(1)$.

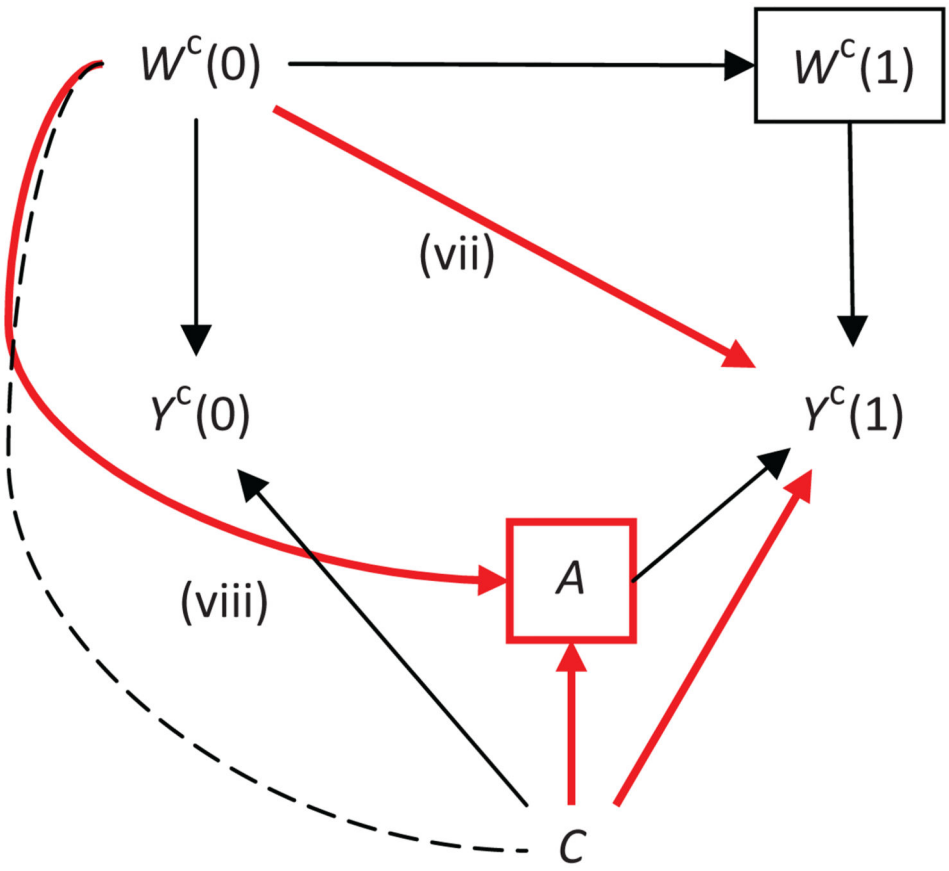

Figure 7.

DAG illustrating the test of independence for the pooled outcome estimand at time t =1 in the presence of unmeasured confounder, C. Two unblocked paths are opened from pre-treatment covariates, $W^c(0)$, to post-treatment outcome, $Y^c(1)$: (vii) through a direct path; and (viii) the path through C created by the collider treatment, A.

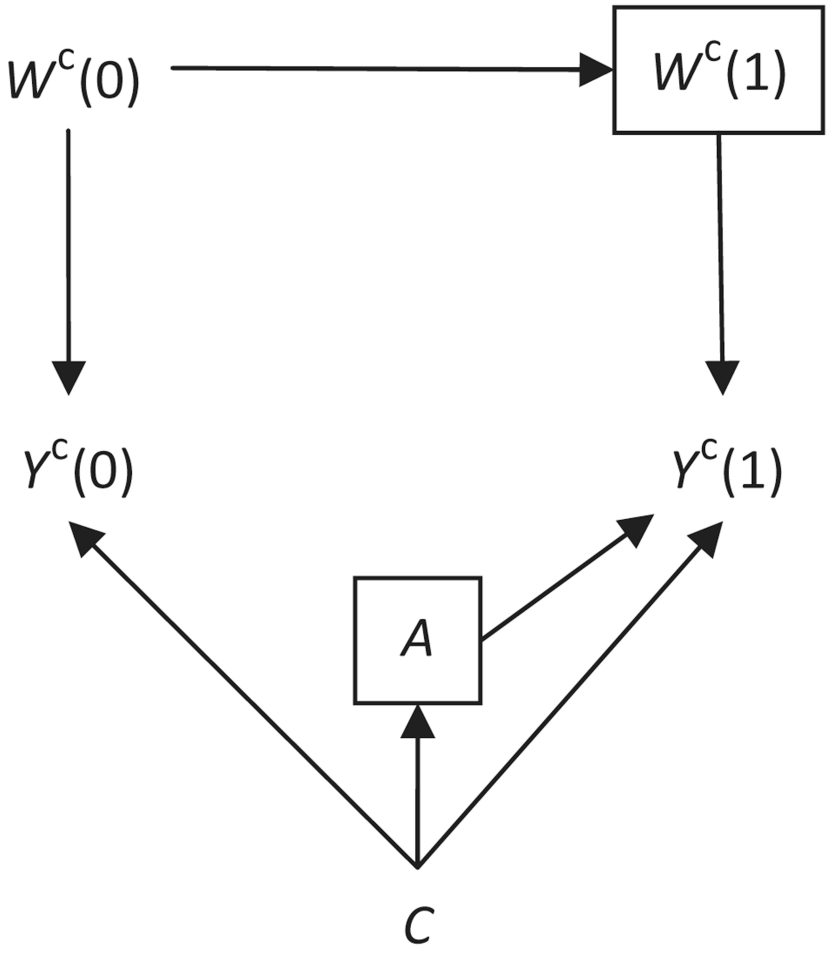

Figure 8.

DAG illustrating exclusion restrictions on pre-treatment covariates $W^c(0)$ for the pooled outcome estimand at time t=1 in the presence of unmeasured confounder, C. The post-treatment outcome, $Y^c(1)$, is independent of $W^c(0)$ given treatment, A, and post-treatment covariates, $W^c(1)$, only if $W^c(0)$ does not affect $Y^c(1)$ and does not affect A.

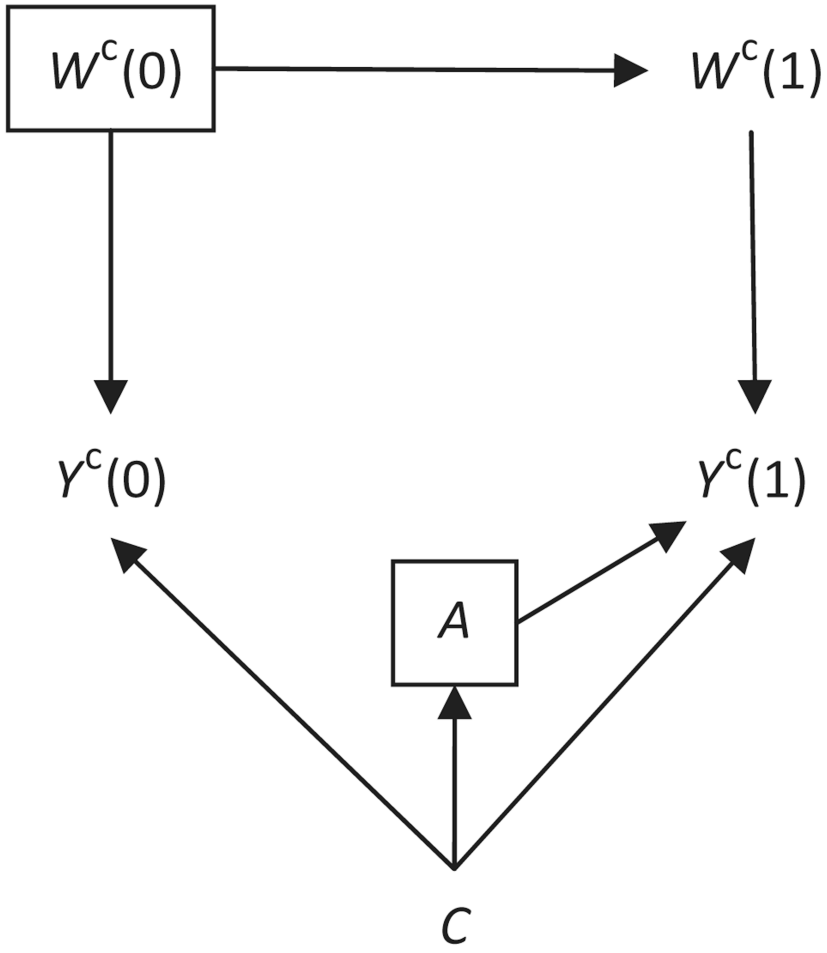

Figure 9.

DAG illustrating exclusion restrictions on pre-treatment covariates $W^c(0)$ for the pooled outcome estimand at time $t=0$ in the presence of unmeasured confounder, C. The lagged outcome, $Y^c(0)$, is independent of post-treatment covariates, $W^c(1)$, given treatment, A, and $W^c(0)$, if $W^c(0)$ does not affect post-treatment outcome, $Y^c(1)$, and does not affect A.

Table 1

Notation used for variables, parameters and outcomes.

Notation	Description
V	Vector of time-invariant community-level covariates† (urban location, province, population size, presence of a health facility, road and water access in wet and dry seasons, indicators of weather shocks between '99 and '01)
$W^c(t)$	Vector of community-level covariates that summarize individual-level factors for the individuals sampled in the community at time t = 0,1 (proportion of mothers sampled in the community who are uneducated or with primary only education, mean child age, proportion of children older than 1 year, proportion of female children, and mean child birth order ranking)
A	Treatment, assigned at the community level
$Y^c(t)=\frac{1}{N}\sum_{i=1}^{N}Y_i(t)$	Community mean of individual-level outcomes $Y_i(t)$ (weight-for-age of child i, i = 1,…,N) for each of the N children under 5 years sampled in the community at time t=0,1
$O_j=\left(V_j, W_j^c(t), A_j, Y_j^c(t)\right)$	Observed data structure, O_j, for a given community j. The observed data are J independent and identically distributed copies of O
$U_V,\ldots U_{Y(t)}$	Sources of random variation for each variable (for example characteristics of leadership in accepting the program (U_A) for A, dispersion of the community across large distances (U_V) for V, lack of a secondary school in the community (U_W) for W, and sampling procedure problems ($U_{Y(t)}$) for $Y(t)$)
P_0	True data-generating distribution; $O_j \sim P_0$
$Y_{\mathrm{a}}^{\mathrm{c}}, Y_{\mathrm{a}}^{\theta}$	Counterfactual outcomes; we focus on two outcomes: the post-treatment outcome, $Y^c(t=1)$, and the change in outcome from pre- to post-treatment, $Y^\theta = Y^c(t=1) - Y^c(t=0)$. For each, we define their counterfactual value under treatment level $A=a$ ($Y_{\mathrm{a}}^{\mathrm{c}}$ and Y_{a}^{θ}, respectively)
$\Psi(P_0)$	True value of the target statistical parameter (or estimand), consisting of parameter mapping Ψ applied to the true data-generating distribution P_0. We present 3 estimands labeled Ψ^{I}, Ψ^{II}, and Ψ^{III}

Note:

†Population size, presence of a health facility and access by road or water information is only available at the community level in 2004. For the purposes of this paper, we assume that these factors did not change significantly from 1997.

Table 2

True value of the estimands[†] under various causal models (true value of target causal parameter=1).

	Simulation[‡] #	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
	Estimand	Ψ^{I}	Ψ^{I}	Ψ^{II}	Ψ^{II}	Ψ^{II}	Ψ^{III}	Ψ^{III}	Ψ^{III}
	True value of estimand	1.0	1.3	1.4	1.0	0.60	1.7	1.0	0.69
Identifying assumptions	*A* does not affect $W^{c}(t=1)$	✓	✓		✓	✓	✓	✓	✓
	$Y^{c}_{a}(t=1) \perp A \mid V, W^{c}(t=0), Y^{c}(t=0), W^{c}(t=1)(\mathrm{RA}(1))$ or $Y^{\theta}_{a} \perp A \mid V, W^{c}(t=0), Y^{c}(t=0), W^{c}(t=1)(\mathrm{RA}(4))$	✓	✓						
	No unmeasured confounder, *C*	✓							
	$Y^{\theta}_{a} \perp A \mid V, W^{c}(t=0), W^{c}(t=1)(\mathrm{RA}(9)$, not conditional on $Y^{c}(t=0))$			✓	✓	✓	✓	✓	✓
	$Y^{c}(t=0)$ does not affect *A*				✓		✓	✓	
	$Y^{c}(t=0)$ does not affect $W^{c}(t=1)$				✓	✓	✓	✓	✓
	$Y^{c}(t=0)$ does not affect $Y^{c}(t=1)$				✓	✓	✓	✓	✓
	$E(Y^{c}(t) \mid A=a, V, W^{c}(t=0), W^{c}(t=1)) = E(Y^{c}(t) \mid A=a, V, W^{c}(t))$ for $t=0, 1$ (Assumption (11))						✓	✓	✓
	$W^{c}(t=0)$ does not affect *A*							✓	✓
	$W^{c}(t=0)$ does not affect $Y^{c}(t=1)$							✓	✓

Notes:

[†]Based on simulation applied to sample of 100,000 observations.

[‡]Reference number for the simulation scenario described in the text.

Table 3

Point treatment effect estimates and confidence intervals using observed data.

Outcome	Estimand	Point estimate	Confidence intervals (LCI, UCI)	Confidence interval method
$Y(t=1)$	Ψ^{I}	0.066	–0.004, 0.135	Normal distribution
			0.001, 0.152	Percentiles
			0.009, 0.123	Influence curve
Y^{θ}	Ψ^{II}	0.255	0.143, 0.368	Normal distribution
			0.140, 0.358	Percentiles
			0.165, 0.346	Influence curve
$Y(t)$	Ψ^{III}	0.274	0.124, 0.425	Normal distribution
			0.232, 0.511	Percentiles
			0.184, 0.365	Influence curve