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

Year 2010 *Paper* 268

Estimation of Causal Effects of Community Based Interventions

Mark J. van der Laan*

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

Estimation of Causal Effects of Community Based Interventions

Mark J. van der Laan

Abstract

Suppose one assigns two interventions to a small number K of different populations or communities, and one measures covariates and outcomes on a random sample of independent individuals from each of the K populations. We investigate the problem of identification and estimation of the causal effect of the choice of intervention assigned at the community level, and, if the intervention is time-dependent, the causal effect of the changes in the intervention at time t, on the outcome. The challenge one is confronted with is that different populations have different environmental factors and that the intervention and environment are assigned to the whole population instead of to the individual. The question we wish to address is if one can still estimate the causal effect of the intervention one would have obtained if one would have combined all units across the multiple populations, each unit having their assigned environment and individual covariates, randomly assign the intervention among the two possible interventions to the unit, and then compare the outcome distributions for the two treatment groups: i.e., if one would have carried out the ideal experiment of randomizing treatment allocation to the units of the combined population, thereby dealing with confounding due to different units having different environments and corresponding individual covariates.

We apply the roadmap based on causal modeling with a nonparametric structural equation model, which involves 1) defining the target causal effect as a parameter on the nonparametric structural equation model, 2) addressing the identifiability from the observed data, and, 3) given an identifiability result under the required assumptions, the efficient estimation of the resulting statistical target parameter through targeted maximum likelihood substitution estimators, using

cross-validation to fine tune the estimators. The fundamental identifiability assumption we make is that one collects baseline covariates on the individual that block the effect of the environment on the outcome of interest, which is formulated as an exclusion restriction assumption in the nonparametric structural equation model.

In addition, we utilize the understanding of the causal identifiability assumptions to evaluate the matched sampling design in which the units of different communities are matched on individual factors. We present efficient weighted targeted maximum likelihood estimators for these matched sampling designs, and we establish the concrete theoretical gain in information for the target parameter relative to independent sampling, by application of general results on case-control biased sampling in van der Laan (2008).

Our methods can be reasonably well applied to the case that the intervention causes infectious behavior among individuals, possibly resulting in an enhanced effect, and to the case that interaction between individuals creates dependence between the individuals. However, the methods would not take into account the effect of this dependence among individuals on the assessment of uncertainty in the point estimates. For that purpose we also propose an estimate of standard error of the point estimate that takes into account arbitrary (and unknown to the user) dependence structures that still permit a central limit theorem based normal approximations.

Our framework and methods are extended to the case that the communities are followed up over time and exposed to a single time-dependent treatment regimen, while also being subjected to changes in environment over time. In particular, we consider the case of estimation of a causal effect of a change in treatment over time based on observing a single community over time under a certain time-dependent treatment regimen.

We also generalize our results to causal effects of combined community based intervention and individually assigned treatment on an outcome of interest. It is shown that G-computation formulas and corresponding estimators developed for causal effects of individually assigned treatments can be fully utilized to estimate these causal effects.

Finally, we consider the case in which one is not willing to assume the exclu-

sion restriction assumption, but many communities are sampled. For that purpose we propose statistical inference that naturally adapts to the degree at which the exclusion restriction assumption is approximated and the number of communities that are sampled. This allows for a unified framework for analyzing studies that involve community based interventions.

1 Introduction.

There is a rich literature on assessment of causal effects of treatment on an outcome based on data at the individual level on a random sample of individuals, both in randomized trials and observational studies. In such studies treatment is "assigned" at the individual level and one also collects covariate and outcome data on each individual. The fundamental problem this part of the causal inference literature needs to address is the utilization of covariates measured at individual level to control for the fact that the treatment empirically or theoretically is a function of such covariates. The fundamental identification problem, involving expressing a well defined causal effect as a parameter of the distribution of the observed data, is addressed by the G-computation formula under the sequential randomization assumption, and, semiparametric model-based efficient estimators of the resulting statistical target parameter of the distribution of the data have been developed.

Over the last years, due to the increasing need to evaluate community based programs in practice, there is growing interest in understanding causal effects of treatments or exposures assigned at the community level, while one still collects data at the individual level for random samples of individuals within these communities. These type of designs result in a new challenge for the semiparametric causal inference literature that seems to not have received much attention yet.

Current practice typically involves using parametric regression models such as mixed linear models. The parametric model approach avoids careful definition of the causal effect of interest as a parameter of the distribution of the data, but instead, one typically focusses on a regression coefficient in a guaranteed misspecified regression model, and one proceeds in estimation, assessment of uncertainty, and interpretation, without acknowledging that the regression model is misspecified. Even if the non-testable causal assumptions for identifiability of a causal effect would be valid, the resulting statistical parameter estimates and inference will be biased.

In this article we aim to study the estimation problem relying on a roadmap involving 1) causal modeling through nonparametric structural equation models (NPSEM), allowing us to define the causal effect of interest, 2) establishing the required identifiability conditions, 3) comitting to a nonparametric/semiparametric statistical models for the data generating distribution implied by these NPSEM, and 4) semiparametric efficient estimation of the statistical target parameter, representing the causal effect under the stated identifiability conditions, with targeted maximum likelihood (substitution) estimators, which naturally integrates the state of the art in machine learning through so called loss-based super learning with semiparametric efficient estimation of a target parameter.

Specifically, suppose one assigns two interventions of interest to K different populations/communities, one takes a random sample of independent units from each of the K populations, and one takes measurements at the individual level on covariates,

and an outcome of interest.

The question we wish to address is, if one can still estimate the effect one would have targeted if one would have combined all units across the multiple populations, sample a unit from this combined population with its environment and individual covariate profile, randomly assign the intervention among the two possible interventions to the unit, and then compare the outcome distributions for the different treatment groups: i.e., if one would have carried out the ideal randomized trial involving random allocation of the treatment choice at the individual/unit level keeping the individual in its environment/neighborhood/community. The challenge one is confronted with is that the treated and control communities have different environmental factors, that the treatment and environment are assigned to the whole population instead of being measured at the individual level, and that only few communities are sampled, thereby not allowing for asymptotics in the number of communities.

We apply the roadmap of causal modeling with a nonparametric structural equation model, defining the target causal effect as a parameter on the nonparametric structural equation model, addressing the identifiability from the observed data under interpretable causal assumptions, and finally the efficient estimation of the resulting target parameter through targeted maximum likelihood estimation, combined with super learning. The fundamental assumption we make is that one collects baseline covariates on the individual that block the effect of the environment on the outcome of interest, which is formulated as an exclusion restriction assumption in the nonparametric structural equation model.

In addition, we utilize our understanding of the causal identifiability assumptions to evaluate the use of matching in the community based studies, thereby aiming to make the different communities similar in their individual covariate distributions. We present efficient weighted targeted maximum likelihood estimators for these matched cohort designs, by application of general results on semiparametric models for case-control biased sampling in van der Laan (2008). This also allows us to evaluate its concrete theoretical gain in information for the target parameter relative to independent random sampling.

Our methods can be reasonably well applied to the case that individuals within a community are correlated, and that the intervention causes infectious behavior among individuals, and thereby possibly an enhanced effect. Even though our NPSEM does not model this enhanced effect, our statistical target parameters implied by the NPSEM will incorporate the enhanced effect. We also provide a general CLT-based method for assessment of uncertainty that takes into account the dependence among individuals, without a need to know underlying independent clusters of units or other type of independence structure.

We will also consider the extreme case in which one observes a single community at the individual level under a single time-dependent treatment regimen. A common problem is to assess a causal effect of a time-dependent intervention strategy

on a particular population that is followed up over time, but, by necessity or by design, only one such time-dependent intervention is carried out. For example, this time-dependent intervention might represent an exposure such as air-pollution for a particular population of individuals, where the airpollution is measured at the community level so that it is the same for each individual.

Off course, it is impossible to nonparametrically identify the causal effect of one treatment regime relative to another treatment regime on the population distribution of a particular outcome of interest, measured at the individual level, from the population distribution of the data, if everybody individual in the target population receives the same treatment. However, such data sets arise naturally and the causal questions are intriguing, and important. The kind of causal questions that arise are questions like "Did the introduction of the death penalty change crime rates?" , "Did the introduction of this particular law (e.g., abortion), change the population distribution of a particular outcome of interest?", "Did the introduction of a marketing television campaign change the population distribution of the behavior of the subjects?", "Did the roll-out of this HIV-prevention program in this country reduce the infection rate?", "Did the change in air-pollution increase or decrease asthma prevalence in California?", "Did the sudden reduction in hormone replacement therapy reduce breast cancer or some other clinical outcome?", "Did the change in greenhouse gasses in the atmosphere cause global warming?", and so on. One observes a change in treatment over time and a change in outcomes of interest, and one wonders if there is a causal relation.

Current practice analyzes these data sets, and these analyses are used to suggest causal effects. Therefore, it is important to provide a formal statistical framework that 1) defines causal parameters as parameters in a nonparametric structural equation model, 2) provides the non-testable assumptions that allow identifiability of the causal target parameters from the observed data distribution, and 3) provide corresponding semiparametric efficient estimators and confidence intervals of the corresponding statistical parameter that represents the causal target parameter under these identifiability conditions.

1.1 Organization of article.

This article is organized as follows. In Section 2 we address the case that one follows up a sample of individuals for each of two communities that is assigned a treatment and control regimen. The identification result and the corresponding proposed targeted maximum likelihood estimators involve adjustment by pre-treatment covariates measured at the individual level in order to block the confounding due to different environments. This assumption will be referred to as an exclusion restriction assumption (on the NPSEM). In particular, the benefit of a matching design is analyzed. The extension of our identifiability results and estimators to the assignment of a time-dependent treatment and control regimen to the two communities is developed as well. In Section 3 we extend our identifiability result and corresponding

targeted ML estimators to K populations.

In Section 4 we consider estimation and inference without assuming the exclusion restriction assumption, thereby operating in a nonparametric statistical model and acknowledging that the statistical target parameter will only approximate the wished causal effect, where the approximation depends on the number of communities and the degree at which the exclusion restriction assumption is violated/holds. Statistical inference w.r.t. the wished causal effect is developed, where the proposed estimate of the standard error data adaptively adapts to the degree of violation of the exclusion restriction assumption and the number of sampled communities.

In Section 5 we consider identification, estimation and inference, of the causal effect of a change in treatment, when one observes a single community exposed to a single time-dependent treatment regimen. In Section 6 we generalize our results to the identification of a causal effect of the community based intervention combined with an individualized assigned treatment. In Section 7 we take a break, and summarize our findings into a practical conclusion. In Section 8 we consider assessment of uncertainty of our proposed estimates that incorporates dependence among observations within the community. We end this article with a summary and some concluding remarks.

1.2 Some literature on statistical methods for analyzing community based interventions.

I acknowledge that I am not very familiar yet with the current literature on causal inference that directly addresses community based interventions. I hope to add relevant references as time processes and welcome suggestions. A helpful article is Oakes (2004), which reviews methods for causal inference for neighborhood effects in social epidemiology. We refer to this article as an overview article putting this causal inference problem in context of the social epidemiology and some of the causal inference literature. Overall, from his article one concludes that the causal inference literature has not focussed much at all on community based interventions (at least up till 2004), and generalized mixed linear regression models, incorporating the hierarchical structure, have dominated this literature instead.

Oakes points out the overlap between the epidemiologists neighborhood effects and educational scientists school effects. The problem addressed by the educational scientists (Raudenbush and Bryk (1986), Raudenbush and Whillms (1995), Raudenbush and Sampson (1999),Coleman et al. (1966), Aitkin and Longford (1986), Goldstein (1995)) is to estimate the effect of teachers on student achievement, while the analogous problem for social epidemiologists is to estimate the effect of toxic dumps, smoking policies, increases in social networking, and so on, on neighborhood's health. He points out that both share the characteristic that the studies are observational and that the data structure is hierarchical in the sense that it involves measurements on both the individuals and the groups in which the individuals operate. So from a statistical point of view there is hardly a difference between these

two schools.

He also points out the long recognized importance of studying contexts such as neigborhoods (Cassel (1976), McMichael (1999), Susser (1999), Berkman et al. (2000), Krieger (2001)). We quote: "Social forces, above and beyond any individual, have been repeatedly shown to play an important role in how we perceive, measure, and address health and illness" (Parsons (1951), Starr (1982), Rose (1985), Clark et al. (1991), Barr (1995), McKinlay (1996), Feldman et al. (1997)). This is the very motivation for the field of social epidemiology, which concerns the study of effects of social forces and relationships on health."

In addition, Oakes states, after having stressed the enormous literature on contextual effects: "Yet due largely to persistent and complex methodological obstacles, along with a lack of attention to them, the causal effect of neighborhood contexts on health continues to confuse and elude us (see Hook (2001)). There appear to be no multilevel neighborhood effect studies with observational data, including those cited above, that directly confront causal inference."

Oakes proceeds to motivates causal models for the mean counterfactual outcome of an individual under set neighborhood interventions, thereby defining a causal effect of a neighborhood intervention on an individual outcome. He presents mixed linear models for the counterfactual mean outcome as a function of an individual and neighborhood specific covariates. He considers the required randomization assumption and experimental treatment assignment assumption, well known in the causal inference literature, under which the coefficients in the mixed linear model can be interpreted as a conditional causal effect, within strata of the covariates that entered the model. He concludes that these non-testable causal assumptions are often unrealistic due to unmeasured confounding and perfectly predictive confounding of the intervention.

Oakes presents the following comment on the enormous use (and abuse) of mixed linear models. We quote from Oakes review "The theoretical foundation of multilevel models lies in variance component methodology, which in its modern form dates back to Fishers work circa 1925 (Draper (1995)). A ground-breaking advance came when Lindley and Smith (1972) formulated their empirical Bayes regression model, but it was not until the introduction of the EM algorithm (Dempster et al. (1977)) that computational feasibility was obtained. Laird and Ware (1982) popularized the model for biostatisticians, Bryk, Raudenbush, Goldstein and Mason for social scientists (Mason et al. (1984), Goldstein (1987), Bryk and Raudenbush (1992)). From our perspective, the widespread (ab)use of the model is due to the recent introduction of user- friendly software, especially HLM and MlWin, and an accessible translation for SAS users by Verbeke and Molenbergs (1997) and Singer (1998). See also Kreft et al. (1994), Leeuw and Kreft (2001). "

Oakes also states: "Understandably, none of the more recent and rigorous discussions of causal inference in either epidemiology or social science (Susser (1973),Greenland (1990), Greenland (2001), Greenland (2002), Manski (1993), Halloran and

Struchiner (1995), Morgenstern (1995), M.E.Sobel (1995), Kaufman and Cooper (1995), Kaufman and Poole (2000), Kaufman and Kaufman (2001), Robins (2001), Maldonado and Greenland (2002)) addressed multilevel neighborhood effects research directly. Finally, none of the many noteworthy general discussions on causal inference with observational data (e.g. Campbell and Stanley (1963), Cochran (1965), McKinlay (1975), Heckman (1979), Leamer (1983), Smith (1990), Rubin (1991), Clogg and Haritou (1997), Copas and Li (1997), Freedman (1997), Winship and Morgan (1999), Pearl (2000), Rosenbaum (2002)) address neighborhood effects or multilevel mod- els, which appear to present some unique issues."

After discussing the lack of identifiability of causal effects from observational studies of neighborhood effects, Oakes proposes randomized community trials as the important way forward. Randomized community trials involve randomly assigning an intervention among a set of possible interventions to a collection of communities/neighborhoods. Clearly, to claim identifiability of a causal effect on an outcome based on such trials, purely based on the fact that the intervention was randomized to a sample community, one will need to sample a large number of communities: i.e., the community now plays the role of the experimental unit in the causal inference literature on observational studies and randomized trials in which treatment is assigned at the individual level.

Examples of randomized community trials, such as mass-media campaigns to improve health knowledge, the repair of bad sidewalks, or community policing initiatives, are provided in (Charlton et al. (1985), Meyer et al. (1991), Shipley et al. (1995), Holder et al. (1997), Feldman et al. (1998), LeFort et al. (1998), Persky et al. (1999), Biglan et al. (2000), Luepker et al. (2000)).

Some of our contributions relative to causal mixed linear models.

In particular, we provide the following contribution to this current literature based on the application of (mixed) linear regression models to assess the causal effect of an intervention assigned to a community. We use NPSEM to define the wished *marginal* causal effect. Initially, we focus on the case that we have large number of observations at the individual level for relatively few communities (e.g., two), so that we cannot rely on asymptotics in the number of communities, and, as a consequence, can also not rely on the intervention being assigned at random to the community. We introduce a concrete interpretable exclusion restriction assumption, namely the existence of individually measured covariates that block the effect of the environment on the outcome, that allows the identification of the causal effect of the community based intervention. Nonetheless, our methods are also extended to the randomized community trials, possibly involving many communities, but in a manner so that the statistical inference will adapt to the degree at which the exclusion restriction assumption holds: if the exclusion restriction holds, the number of individuals will drive the precision, and if it fails to hold, the number of communities will drive the precision, and in the more realistic grey zone, it adapts naturally.

We assume nonparametric or semi-parametric models for the observed data distribution, thereby avoiding the bias in effect estimates and statistical inference due to miss-specified linear regression models. This semi-parametric model approach requires defining the statistical target parameter (i.e., causal parameter of interest under the causal assumptions) as a parameter of the observed data distribution for any possible observed data distribution, avoiding the common misplaced practice of defining an effect as a coefficient in a misspecified parametric model.

To deal with confounding, matching by design plays an important role, beyond the statistical adjustment. For that purpose, we incorporate matched cohort designs in which the communities receiving the treatment are matched with the communities receiving the control w.r.t. their covariate distributions for the individuals, for a few of the measured covariates. For each of the statistical estimation problems we develop the efficient targeted maximum likelihood methodology to obtain the wished causal effect estimates and corresponding confidence intervals. We also generalize our results to arbitrary causal parameters of interest of interventions that have a community component and an individually assigned treatment component.

2 Assigning two interventions to two different communities.

We consider a study that involves assigning a treatment to one population and a control treatment to another population, sampling units from the treatment and control population, and measuring covariates and outcome on each sampled unit.

Let $A \in \{a_0, a_1\}$ be a variable indicating the two treatment-regimens, and let $E \in \{e_0, e_1\}$ be a variable indicating the two regimens of environmental factors that supposedly measures well the differences in environment relevant to the outcome of interest. For convenience, we will often refer to $A = 1$ for $A = a_1$ and $A = 0$ for $A = a_0$, and we will refer to $A = 0$ as control. For the population that is exposed to treatment we have $(A, E) = (1, e_1)$ and for the population that is exposed to control we have $(A, E) = (0, e_0)$. Clearly, the experimenter that is interested in assessing the causal effect of $A = 1$ versus $A = 0$ would prefer to see that $e_0 = e_1$. The treatment could be a time-dependent exposure over a time-window.

In this section we first focus on establishing the causal effect of assigning the whole regimen $A = 0$ (i.e. $A = (a_0(t) : t)$) versus $A = 1$ (i.e. $A = (a_1(t) : t)$), while later we will also address estimation of the causal effect of just the t-specific component $A(t) = a_0(t)$ versus $A(t) = a_1(t)$ at time t.

Typically, the realization of (A, E) is generated as follows: one would select two populations, whose environment defines two e-profiles, and then one randomly assigns the two possible treatments to these two regions, giving the realized $(1, e_1)$, $(0, e_0)$ for (A, E). The combined population represents the target population of units, and our causal NPSEM below describes a mechanism for assigning $(E, A) \in \{(e_1, 1), (e_1, 0), (e_0, 1), (e_0, 0)\}$ to a sampled unit from that target population, and

subsequently measuring covariate and outcome data on that unit. This NPSEM allows us to define the outcome distribution under set treatment, keeping the selection of the environment random, and define corresponding causal effects of treatment on the outcome.

Observed data. For the treated population we sample n_1 units providing n_1 i.i.d observations of (M_1, Y_1) defined as a draw from the conditional probability distributions of (M, Y), given $(A, E) = (1, e_1)$. Similarly, we sample n_0 units from the control population providing n_0 i.i.d observations of (M_0, Y_0) from the conditional distribution of (M, Y), given $(A, E) = (0, e_0)$. These random variables M, Y could be time-dependent processes.

2.1 The causal model and causal parameter.

In this subsection we formally define the causal effect by a NPSEM, and provide the link to the observed data, laying the ground work for addressing the identifiability from the observed data.

NPSEM. Let $M = (W = M(0), M(1))$ and assume that W are measurements on the unit taken before it was exposed to treatment. For example, the time ordering for the variables measured on the unit might be given by

$$E - W - A - M(1) - Y.$$

What matters is that we know that W is only a function of E, and not of A. Since our target population of units is the combined population, we assume that E has only two possible values e_0, e_1, so that E is also a binary variable. The NPSEM with endogenous variables $X = (E, W, A, M(1), Y)$ is given by:

$$\begin{aligned} U &= (U_E, U_W, U_A, U_{M(1)}, U_Y) \sim P_U \\ E &= f_E(U_E) \\ W &= f_W(E, U_W) \\ A &= f_A(E, W, U_A) \\ M(1) &= f_{M(1)}(E, W, A, U_{M(1)}) \\ Y &= f_Y(E, W, A, M(1), U_Y). \end{aligned}$$

This defines a random variable (U, X) on the unit. This NPSEM allows us to define counterfactuals such as $Y(1) = Y(A = 1)$, $Y(0) = Y(A = 0)$ corresponding with setting $A = 1$ and $A = 0$, respectively. Similarly, it defines random variables $Y(e, a)$ corresponding with interventions setting $E = e, A = a$. These counterfactuals are random variables defined as functions of (U, X) obtained by intervening on the system that generates (U, X).

Link to observed data. We will assume that U_A, U_E are independent of $U_W, U_{M(1)}, U_Y$: i.e., we assume that A, E are randomized, which is a natural assumption on the NPSEM. In that case, the observed data can be viewed as two i.i.d. samples of size n_1, n_0 from the counterfactual (post-intervention) distribution for the intervention $A = 1, E = e_1$ and $A = 0, E = e_0$, respectively. That is, one observes n_1 i.i.d observations on the counterfactual $(W(e_1), M(1)(a_1, e_1), Y(a_1, e_1))$ and n_0 i.i.d. observations on the counterfactual $(W(e_0), M(1)(a_0, e_0), Y(a_0, e_0))$.

Our goal is now to define the parameter of interest/causal effect of interest on the NPSEM, as a parameter of the distribution of (U, X), and then, under certain additional assumptions on the NPSEM, establish identifiability of this causal effect from the two post intervention distributions P_{1,e_1} and P_{0,e_0} identified by our observed data.

Target parameter on NPSEM: We define our parameter of interest in the NPSEM for $X = (E, W, A, M(1), Y)$ as

$$\Psi^F(P_{U,X}) = E\{Y(1) - Y(0)\}, \tag{1}$$

where the reader is reminded that $Y(a)$ is the counterfactual defined by setting $A = a$, for $a \in \{0, 1\}$. This additive causal effect of A on Y corresponds with randomly assigning treatment or control to each unit in the combined population that has characteristics measured by (E, W), and taking the difference in means for the treatment and control group. Such an ideal experiment would thus create a treatment group and control group that has units with both e_0 and e_1-environments, and these environmental factors would be approximately balanced between the treatment and control group. That is, this target parameter is free from environmental confounding.

2.2 Identifiability of target causal effect from observed data.

We now need to address the identifiability of $E\{Y(1) - Y(0)\}$ from the probability distributions P_{1,e_1} and P_{0,e_0} from which we have two samples. For this purpose we make the additional assumption on the NPSEM that the effect of E on Y only goes through W: i.e, the f_Y equation in the above NPSEM is replaced by

$$Y = f_Y(W, A, M(1), U_Y). \tag{2}$$

Under this exclusion restriction assumption (2) and the strong randomization assumption stating that (E, W, A) is independent of $Y(e, w, a)$ in the NPSEM, we have the following identifiability result:

$$EY(1) - EY(0)$$
$$= \sum_w [E(Y(1, e_1) \mid W(1, e_1) = w) - E(Y(0, e_0) \mid W(0, e_0) = w)]P(W = w),$$

where

$$\begin{aligned} P(W=w) &= P(E=e_1)P(W=w \mid E=e_1) + P(E=e_0)P(W=w \mid E=e_0) \\ &= \alpha P_{1,e_1}(w) + (1-\alpha)P_{0,e_0}(w), \end{aligned}$$

and $\alpha = P(E=e_1)$. We also assume that α is known, or equivalently, that the marginal distribution of E in the NPSEM is known. A standard choice for α is given by $\alpha = n_1/n$, which corresponds with a combined population in which the population is weighted by the number of observations sampled from that population (which might be proportional to the population size of that population). Thus, α can be viewed as a choice that defines the target combined population on which we wish to define the additive causal effect of setting treatment versus control.

Heuristics behind "no residual environmental confounding" assumption (2). The idea behind this assumption (2) is that e_1 (e_0), although common to all units in the region, results in unit specific effects of e_1 (e_0) on Y, which is some function $f()$ of characteristics C of the unit and e_1 (e_0). Suppose we are able to observe this particular function of the characteristics C of the unit and e_1 for each unit, so that it is captured by W: e.g., $W(e_1) = f(C_1, e_1)$ is this particular function of e_1 and the characteristics. Similarly, $W(e_0) = f(C_0, e_0)$ is this same function of e_0 and the characteristics of the unit in the control population. By controlling for $W = W(E)$, we are then able to control for the difference in environments for the two populations (i.e., $e_0 \neq e_1$) at the individual level. Even if W does not succeed in capturing the complete effect of e on the unit-specific outcome Y, controlling for it, will still help to take away some of the difference in outcome distributions of Y that is purely due to the differences between the two environmental profiles e_0 and e_1.

Let's consider an example. Consider a study that is interested in evaluating the causal effect of an intervention such as circumcision/diaphram/condom use/ in preventing HIV-infection. For that purpose we consider two cohorts of non-infected individuals from two different regions, and in one region everybody gets exposed to the intervention (e.g., all circumcised, exposed to educational program, and so on), and the other cohort is a control region. The outcome measured at the individual level is the HIV-infection status during the course of the study. Comparing the infection rates in the two cohorts is problematic since it is known that the proportion of HIV-infected individuals in the treatment region is different from the proportion of HIV-infected individuals in the control region. This is an example of different environments that will affect the chance of an individual to become infected: i.e, there is a higher probability of being infected in one cohort versus the other cohort. What covariates should we measure to block this effect of the different region-specific infection rates? Let the proportion of infected individuals for the treatment and control region be r_1 and r_0, respectively. Suppose we measure at the individual level the average number of sexual contacts and partners per month. We could now

propose an individual risk measure for being infected with HIV at baseline (i.e., before the treatment starts): e.g., for an individual in the treatment region it might be r_1 times number of sexual partners, and for an individual in the control region it is r_0 times the number of sexual partners. We include this covariate as a component of W, giving us a component of $W(e_1)$ and $W(e_0)$. So, given two people with similar sexual lifestyles, the person in a low risk environment will get a lower level assigned to this risk measure than the person in a high risk environment. One might now argue that this covariate will help to block the effect of the differential infection rates in the two regions, and thereby makes the exclusion restriction assumption more reasonable: one might argue that a person in the treated region and control region with the same value for this blocking covariate and other pre-treatment covariates are now having the same probability of being infected during this study.

We will state the identifiability result formally as a theorem. For the sake of simplicity, we will ignore the intermediate covariate $M(1)$ since it plays no role in the identifiability result.

Theorem 1 NPSEM. *Consider a NPSEM with structural equations for the endogenous $X = (E, W, A, Y)$,*

$$\begin{aligned} E &= f_E(U_E) \\ W &= f_W(E, U_W) \\ A &= f_A(E, W, U_A) \\ Y &= f_Y(E, W, A, U_Y), \end{aligned}$$

and exogenous $U = (U_E, U_W, U_A, U_Y)$. Let $A \in \{0,1\}$, $E \in \{e_0, e_1\}$ and let $\alpha = P(E = e_1)$ be known.

Counterfactuals. *Let $Y(1) = f_Y(W, 1, U_Y)$ and $Y(0) = f_Y(W, 0, U_Y)$ denote the counterfactuals corresponding with setting $A = 1$ and $A = 0$, respectively. We also define $(W(e_1), Y(1, e_1))$ and $(W(e_0), Y(0, e_0))$ as the post-intervention random variable corresponding with setting $A = 1, E = e_1$ and $A = 0, E = e_0$, respectively. We also define $Y(e, w, a) = f_Y(w, a, U_Y)$ as the post-intervention counterfactual of Y corresponding with intervention $E = e, W = w, A = a$. We denote the distributions of $(W(e_1), Y(1, e_1))$ and $(W(e_0), Y(0, e_0))$ with P_{1,e_1} and P_{0,e_0}, respectively.*

Observed data. *Let $O = (B, W(B) \equiv W(e_B), Y(B))$, where $B \sim$ Bernoulli$(\alpha) \in \{(0, e_0), (1, e_1)\}$, conditional on $B = (1, e_1)$, O is distributed as $(W(e_1), Y(1, e_1)) \sim P_{1,e_1}$, and, conditional on $B = (0, e_0)$, O is distributed as $(W(e_0), Y(0, e_0)) \sim P_{0,e_0}$. In particular, we note that the marginal distribution of B equals the marginal distribution of E. We also note that $P(W(B) = w) = P_{W(e_1)}(w)\alpha + P_{W(e_0)}(w)(1-\alpha)$. Let P_O be the probability distribution of O: $P_O = P_O(P_{U,X})$. We observe n i.i.d. observations on O.*

Relevance to two sample problem. *We note that the distribution of P_O also approximates the two sample experiment in which one samples n_0 i.i.d. observations from P_{0,e_0}, and n_1 i.i.d. observations from P_{1,e_1}, with $n_1/(n_0 + n_1) = \alpha$.*

Target parameter on NPSEM. *Consider the following parameter of the distribution of* (U, X)*:*

$$\Psi^F(P_{U,X}) = EY(1) - EY(0).$$

Exclusion and Randomization assumption on NPSEM. *Assume that* Y *is only a function of* E *through* W*, i.e.,* $Y = f_Y(W, A, U_Y)$ *in the NPSEM. Assume also that the distribution of* $U = (U_E, U_W, U_A, U_Y)$ *is such that* (E, W, A) *is independent of* $Y(e, w, a)$ *for all* e, w, a*.*

Identifiability Result. *Then*

$$\begin{aligned}\Psi^F(P_{U,X}) &= E_{W(B)}\{E(Y(B) \mid W(B), B = (1, e_1)) - E(Y(B) \mid W(B), B = (0, e_0))\} \\ &\equiv \Psi(P_0).\end{aligned}$$

Proof. Firstly, for the full data parameter, we have

$$\begin{aligned}&\psi_0^F = EY(1) - EY(0) \\ &= Ef_Y(1, W, U_Y) - Ef_Y(0, W, U_Y) \\ &= \textstyle\sum_w \{E(f_Y(1, w, U_Y) \mid W = w) - E(f_Y(0, w, U_Y) \mid W = w)\} P(W = w) \\ &= \textstyle\sum_w \{Ef_Y(1, w, U_Y) - Ef_Y(0, w, U_Y)\} P(W = w),\end{aligned}$$

where we used that W is independent of $Y(e, w, a)$, by assumption. We note that $P(W = w) = P(W(e_1) = w \mid E = e_1)P(E = e_1) + P(W(e_0) = w \mid E = e_0)P(E = e_0) = P_{W(e_1)}(w)\alpha + P_{W(e_0)}(w)(1 - \alpha)$.

Consider now the parameter ψ_0 of observed data. Since, given $B = (1, e_1)$, $(W(B), Y(B))$ is distributed as $(W(e_1), Y(1, e_1))$, we have

$$\begin{aligned}E(Y(B) \mid W(B) = w, B = (1, e_1)) &= E(f_Y(1, w, U_Y) \mid W(e_1) = w) \\ &= E(f_Y(1, w, U_Y) \mid A = 1, E = e_1, W = w) \\ &= Ef_Y(1, w, U_Y),\end{aligned}$$

where the second equality is implied by (A, E) being independent of $Y(e, w, a)$, given W, and the third equality is implied by (E, W, A) being independent of $Y(e, w, a)$, both consequences of our strong randomization assumption.

Similarly, $E(Y(B) \mid W(B) = w, B = (0, e_0)) = Ef_Y(A = 0, w, U_Y)$. In addition, ψ_0 involves averaging w.r.t $P(W(B) = w) = P_{W(e_1)}(w)\alpha + P_{W(e_0)}(w)(1 - \alpha)$.

Thus,

$$\psi_0 = \sum_w \{Ef_Y(A1, w, U_Y) - Ef_Y(A = 0, w, U_Y)\} P(W = w),$$

which is thus identical to ψ_0^F. This completes the proof. □

Commitment to statistical parameter and model for observed data. Based

on this theorem, we propose the following parameter of the distribution of the observed data structure $O = (B, W(B), Y(B))$ with B bernoulli in $\{(1, e_1), (0, e_0)\}$:

$$\Psi(P_O) \equiv E_{W(B)} \{E(Y(B) \mid W(B), B = (1, e_1)) - E(Y(B) \mid W(B), B = (0, e_0))\}.$$

Under the NPSEM, the marginal distribution of E being known, the strong randomization assumption that (E, W, A) is independent of $Y(e, w, a)$ as well as the assumption that E affects Y only through W, as stated in Theorem, we have that $\Psi(P_O) = EY(1) - Y(0)$. Either way, we suggest that $\Psi(P_0)$ is also an interesting treatment effect measure as a pure statistical parameter, i.e., without the causal assumptions, but its causal interpretation under these non-testable assumptions, adds a lot of flavor to this statistical parameter.

The model for the probability distribution P_0 of $O = (B, W = W(B), Y = Y(B))$ is nonparametric, and the statistical target parameter is $\Psi(P_O) = E_W E(Y \mid B = (1, e_1), W) - E(Y \mid B = (0, e_0), W)$.

2.3 Estimation and inference.

The targeted maximum likelihood estimator of this statistical parameter has been defined previously and statistical inference as well (see, e.g., van der Laan and Rubin (2006) for the targeted MLE, and van der Laan and Gruber (2010), Gruber and van der Laan (2010) for the collaborative targeted MLE). Since we have arrived at the pure statistical estimation stage, we will denote $O = (B, W = W(B), Y = Y(B))$ and the pooled sample with O_i, $i = 1, \ldots, n = n_1 + n_2$. One starts out with applying a super learner to fit $Q_0 = E(Y \mid B, W)$ and subsequently one applies targeted maximum likelihood estimation to update this initial super learner estimate Q_n^0. The marginal distribution of W is estimated with the empirical of the pooled sample W_i, $i = 1, \ldots, n = n_1 + n_2$. The targeted maximum likelihood estimate requires a fit of $P(B = 1 \mid W)$. The estimator is double robust in the sense that it remains unbiased if one either consistently estimates $g_0(1 \mid W) \equiv P(B = 1 \mid W)$ or $Q_0(B, W) \equiv E(Y \mid B, W)$. The estimator is efficient if the initial estimator Q_n^0 is consistent and g_n is consistent as well, and if g_n is misspecified (but Q_n^0 is consistent), it can both be super efficient as well as inefficient, depending on its limit. The targeted maximum likelihood estimator can be further refined with the collaborative targeted maximum likelihood estimation method, resulting in a collaborative double robust estimator, that has generally better finite sample efficiency, and is consistent under weaker conditions.

This double robustness of the targeted maximum likelihood estimator in terms of the factors g_0, Q_0 of the distribution of (W, B, Y), translates into the following robustness in terms of the distributions P_{0,e_0}, P_{1,e_1} for the two samples. Firstly, we note that $\bar{P}(w) \equiv P(W = w) = \alpha P_{W(e_1)}(w) + (1 - \alpha) P_{W(e_0)}(w)$. We also define $Q_1(w) = E(Y(e_1, 1) \mid W(e_1) = w)$, $Q_0(w) = E(Y(e_0, 0) \mid W(e_0) = w)$, and we note

that

$$\begin{aligned} Q_0(B,W) &= E(Y \mid B,W) = I(B=1)Q_1(W) + I(B=0)Q_0(W), \\ g_0(1 \mid W) &= P(B=1|W=w) = \frac{P(B=1,W=w)}{P(W=w)} = \frac{P(W=w|B=1)P(B=1)}{P(W=w)} \\ &= \frac{P_{e_1}(w)\alpha}{\bar{P}(w)}, \\ g_0(0 \mid W) &= P(B=0|W=w) = \frac{P_{e_0}(w)(1-\alpha)}{\bar{P}(w)}. \end{aligned}$$

Thus, the double robustness of the targeted MLE for estimation of ψ_0 in a nonparametric model for $O = (B, W_B, Y_B)$ in terms of g_0, Q_0 can be restated as follows: the targeted MLE will be consistent if either the outcome regressions Q_1, Q_0 on the covariates are consistently estimated for both samples, or if the ratio P_1/P_0 of the covariate distributions for the two samples is correctly estimated. In particular, the identifiability condition $0 < P(B = 1 \mid W) < 1$ a.e. translates into $0 < \alpha < 1$, and that the Radon-Nykodym derivatives $P_{e_0}(w)/P_{e_1}(w) < \infty$ and $P_{e_1}(w)/P_{e_0}(w) < \infty$ for the covariate distributions are bounded. Thus, if a covariate can have a certain value in population 1, then that value should also occur in population 2, and visa versa.

Statistical inference for the targeted MLE can be based on the influence curve for ψ_0 in the nonparametric model for $O = (B, W = W(B), Y = Y(B))$, given by

$$D^*(O) = \frac{1-2B}{g_0(B \mid W)}(Y - Q_0(B,W)) + Q_0(1,W) - Q_0(0,W) - \psi_0.$$

That is, one can estimate the asymptotic variance of the targeted MLE with $\sigma_n^2 = 1/n \sum_{i=1}^n \hat{D}^{*2}(O_i)$ and an asymptotic 0.95-confidence interval for ψ_0 is given by $\psi_n \pm 1.96\sigma_n/\sqrt{n}$, where $\hat{D}$ is the estimate of the efficient influence curve obtained by substituting the estimates g_n, Q_n of g_0, Q_0.

2.4 Causal effect among the treated population.

We will consider another statistical parameter $\psi_0^t = E_W\{E(Y \mid W, B = (1, e_1)) - E(Y \mid W, B = (0, e_0))\}$, which is referred to as the treatment effect among the treated. Even though our results regarding the matched cohort design only concern the statistical target parameter, for the sake of interpretation, we like to know its causal interpretation under the NPSEM, the exclusion, and strong randomization assumption. The following theorem provides us with the wished identifiability result.

Theorem 2 NPSEM. *Consider a NPSEM with structural equations for the en-*

dogenous $X = (E, W, A, Y)$,

$$\begin{aligned} E &= f_E(U_E) \\ W &= f_W(E, U_W) \\ A &= f_A(E, W, U_A) \\ Y &= f_Y(E, W, A, U_Y), \end{aligned}$$

and exogenous $U = (U_E, U_W, U_A, U_Y)$. *Let* $A \in \{0, 1\}$, $E \in \{e_0, e_1\}$ *and let* $\alpha = P(E = e_1)$.

Counterfactuals. *Let* $Y(1) = f_Y(W, 1, U_Y)$ *and* $Y(0) = f_Y(W, 0, U_Y)$ *denote the counterfactuals corresponding with setting* $A = 1$ *and* $A = 0$, *respectively. We also define* $(W(e_1), Y(1, e_1))$ *and* $(W(e_0), Y(0, e_0))$ *as the post-intervention random variable corresponding with setting* $A = 1, E = e_1$ *and* $A = 0, E = e_0$, *respectively. We also define* $Y(e, w, a) = f_Y(w, a, U_Y)$ *as the post-intervention counterfactual of* Y *corresponding with intervention* $E = e, W = w, A = a$. *We denote the distributions of* $(W(e_1), Y(1, e_1))$ *and* $(W(e_0), Y(0, e_0))$ *with* P_{1,e_1} *and* P_{0,e_0}, *respectively.*

Observed data. *Let* $O = (B, W(B) \equiv W(e_B), Y(B))$, *where* $B \sim$ *Bernoulli*$(\alpha) \in \{(0, e_0), (1, e_1)\}$, *conditional on* $B = (1, e_1)$, O *is distributed as* $(W(e_1), Y(1, e_1)) \sim P_{1,e_1}$, *and, conditional on* $B = (0, e_0)$, O *is distributed as* $(W(e_0), Y(0, e_0)) \sim P_{0,e_0}$. *In particular, we note that the marginal distribution of* B *equals the marginal distribution of* E. *We also note that* $P(W(B) = w) = P_{W(e_1)}(w)\alpha + P_{W(e_0)}(w)(1 - \alpha)$. *Let* P_O *be the probability distribution of* O: $P_O = P_O(P_{U,X})$. *We observe* n *i.i.d. observations of* O.

Relevance to two sample problem. *We note that a sample of* $n = n_0 + n_1$ *i.i.d. observations from the distribution of* P_O *also approximates the two sample experiment in which one samples* n_0 *i.i.d. observations from* P_{0,e_0}, *and* n_1 *i.i.d. observations from* P_{1,e_1}, *in which case* $n_1/(n_0 + n_1) = \alpha$.

Target parameter on NPSEM. *Consider the following parameter of the distribution of* (U, X):

$$\Psi^F(P_{U,X}) = E[Y(1) - Y(0) \mid (A, E) = (1, e_1)).$$

Exclusion and Randomization assumption on NPSEM. *Assume that* Y *is only a function of* E *through* W, *i.e.,* $Y = f_Y(W, A, U_Y)$ *in the NPSEM, and that the distribution of* $U = (U_E, U_W, U_A, U_Y)$ *is such that* (E, W, A) *is independent of* $Y(e, w, a)$ *for all* e, w, a. *We also assume that the distribution of* E *is known.*

Identifiability result. *Then*

$$\begin{aligned} &\Psi^F(P_{U,X}) = \Psi^t(P_O) \\ &\equiv \textstyle\sum_w P_{e_1}(w))\{E(Y(B) \mid W(B) = w, B = (1, e_1)) - E(Y(B) \mid W(B) - w, B = (0, e_0))\}, \end{aligned}$$

where $P_{e_1}(w) = P(W(e_1) = w) = P(W(B) = w \mid B = (1, e_1))$.

Proof. Firstly, for the full data parameter, we have

$$\begin{array}{l}\psi_0^F = E[Y(1) - EY(0) \mid (E,A) = (e_1,1))\\ = E[f_Y(1,W,U_Y) \mid (E,A) = (e_1,1)] - E[f_Y(0,W,U_Y) \mid (E,A) = (e_1,1)]\\ = \sum_w E[f_Y(1,w,U_Y) \mid W = w, (E,A) = (e_1,1))P(W(e_1) = w)\\ - \sum_w E(f_Y(0,w,U_Y) \mid W = w, (E,A) = (e_1,1))P(W(e_1) = w)\\ = \sum_w \{Ef_Y(1,w,U_Y) - Ef_Y(0,w,U_Y)\}P(W(e_1) = w),\end{array}$$

where we used that E, W, A is independent of $Y(e,w,a)$, by assumption.

Consider now the parameter ψ_0^t of the distribution of the observed data. Since, given $B = (1,e_1)$, $(W(B), Y(B))$ is distributed as $(W(e_1), Y(1,e_1))$, we have

$$\begin{array}{rcl}E(Y(B) \mid W(B) = w, B = (1,e_1)) & = & E(f_Y(1,w,U_Y) \mid W(e_1) = w)\\ & = & E(f_Y(1,w,U_Y) \mid A = 1, E = e_1, W = w)\\ Ef_Y(1,w,U_Y), & & \end{array}$$

where the second equality is implied by (A,E) being independent of $Y(e,w,a)$, given W, and the third equality is implied by (E,W,A) being independent of $Y(e,w,a)$, both consequences of our strong randomization assumption. Similarly, it follows that $E(Y(B) \mid W(B) = w, B = (0,e_0)) = Ef_Y(A = 0,w,U_Y)$. In addition, ψ_0 involves averaging w.r.t $P(W(e_1) = w)$.

Thus,

$$\psi_0 = \sum_w \{Ef_Y(A1,w,U_Y) - Ef_Y(A = 0,w,U_Y)\}P(W(e_1) = w),$$

which is identical to ψ_0^F. This completes the proof. □

2.5 The causal effect among the treated, and its targeted MLE.

In this subsection we consider the statistical parameter $\Psi^t(P_0) = E_{W(e_1)}\{E(Y(B) \mid B = (1,e_1), W(B) = W(e_1)) - E(Y(B) \mid B = (0,e_0), W(B) = W(e_1))\}$ defined above, an effect among the treated. This parameter will play an important role in the next subsections, since matched cohort design will be shown to be particularly optimal for targeting this effect among the treated. In this subsection we will develop the targeted MLE for this parameter.

Suppose we observe n i.i.d. observations of $O = (W,A,Y)$, W baseline covariates, subsequently assigned binary treatment A, and final outcome Y of interest. Note, in our application, we have $O = (W,B,Y)$ so that B will play the role of A in the sequel of this subsection. The statistical parameter of interest is then given by $E(E(Y \mid A = 1, W) - E(Y \mid A = 0, W) \mid A = 1)$.

Suppose the model is nonparametric and we wish to estimate the following parameter of the data generating distribution P_0 of $O = (W,A,Y)$

$$\Psi(P_0) = E_0\left\{E_0(Y \mid A = 1, W) - E_0(Y \mid A = 0, W) \mid A = 0\right\}.$$

Under an NPSEM implied by the ordering W, A, Y and the randomization assumption $A \perp Y(a)$, given W, one can interpret this parameter as $E(Y(1)-Y(0) \mid A=0)$.

Another way of representing this parameter is $\Psi(P_0) = -E_0(Y - E(Y \mid A = 1, W) \mid A = 0)$, i.e., among the non-treated one evaluates the outcome minus the predicted outcome if, contrary to the fact, one would have been treated, and one takes the population average of all these differences.

Suppose one wishes to estimate the effect among the treated, as in our application, given by

$$\Psi_1(P_0) = E_0\left\{E_0(Y \mid A=1, W) - E_0(Y \mid A=0, W) \mid A=1\right\},$$

which under the causal assumptions can be represented as $E(Y(1) - Y(0) \mid A = 1)$. Switching the roles of $A = 1$ and $A = 0$ in the formulas below provides the efficient influence curve and targeted MLE of $-\Psi_1(P_0)$. We will make this specific below.

Note that P_0 is determined by the marginal distribution P_W of W, the conditional distribution $P_{A|W}$ of A, given W, and the conditional distribution $P_{Y|A,W}$ of Y, given A, W. The parameter $\Psi(P_0)$ depends on P_0 through both $P_W, P_{Y|A,W}$ as well as the treatment mechanism $P_{A|W}$. We will denote the treatment mechanism with g_0 and the other two factors of the likelihood with Q_0.

The efficient influence curve of the target parameter. Firstly, consider the parameter $P \to \Psi(P)(1) = E_P(E_P(Y \mid A = 1, W) \mid A = 0)$. The efficient influence curve of this parameter is given by

$$D_1^*(P) = \frac{I(A=1)}{P(A=0)}\frac{g(0 \mid W)}{g(1 \mid W)}(Y - Q(1, W)) + \frac{I(A=0)}{P(A=0)}(Q(1, W) - \Psi(P)(1)).$$

Here $Q(P)(a, W) = E_P(Y \mid A = a, W)$ and $g(P)(a \mid W) = P(A = a \mid W)$.

The efficient influence curve of $\Psi(P)(0) = E_P(E_P(Y \mid A = 0, W) \mid A = 0)$ at P is given by

$$D_0^*(P) = \frac{I(A=0)}{P(A=0)}(Y - Q(0, W)) + \frac{I(A=0)}{P(A=0)}(Q(0, W) - \Psi(P)(0)).$$

Thus the efficient influence curve of $\Psi(P) = \Psi(P)(1) - \Psi(P)(0)$ is given by

$$\begin{aligned} D^*(P) = & \left\{\frac{I(A=1)}{P(A=0)}\frac{g(0 \mid W)}{g(1 \mid W)} - \frac{I(A=0)}{P(A=0)}\right\}(Y - Q(A, W)) \\ & + \frac{I(A=0)}{P(A=0)}\left\{Q(1, W) - Q(0, W) - \Psi(P)\right\}. \end{aligned}$$

The efficient influence curve of $\Psi(P) = E_P\{E_P(Y \mid A = 1, W) - E_P(Y \mid A = 0, W) \mid A = 1\}$ is obtained by changing roles of $A = 1$ and $A = 0$, and taking minus

sign, giving

$$\begin{aligned} D^*(P) &= \left\{ \frac{I(A=1)}{P(A=1)} - \frac{I(A=0)}{P(A=1)} \frac{g(1 \mid W)}{g(0 \mid W)} \right\} (Y - Q(A,W)) \\ &\quad + \frac{I(A=1)}{P(A=1)} (Q(1,W) - Q(0,W) - \Psi(P)). \end{aligned}$$

Double robustness of efficient influence curve. This efficient influence curve of $\Psi(P)$ can be represented as an estimating function $D^*(Q, g, \psi)$, where we suppress the dependence on the scalar $P(A = 0)$. We note that this estimating function is double robust in the sense that it is an unbiased estimating function for ψ_0, if either Q is correctly specified, or g is correctly specified. Formally, this is stated as

$$P_0 D^*(Q, g, \psi_0) = 0 \text{ if } Q = Q_0 \text{ or } g = g_0,$$

and $g(1 \mid W) > 0$ a.e. Here we recall the notation $Pf \equiv \int f(o) dP(o)$. This double robustness result can be explicitly verified.

In fact, we can establish a stronger so called collaborative double robustness, defined as follows. Let $W(Q)$ be a subset/reduction of W so that conditioning on $W(Q)$ also fixes $(Q - Q_0)(a, W)$ for $a \in \{0, 1\}$. Then, for all Q and corresponding $g_0(Q) = P(A = \cdot \mid W(Q))$ for such a $W(Q) \subset W$, we have

$$P_0 D^*(Q, g_0(Q), \psi_0) = 0.$$

Note that this implies, in particular, $P_0 D^*(Q_0, g) = 0$ for all g, since, if $Q = Q_0$, then we can select $W(Q)$ as the empty set. Thus, g_0 only needs to adjust for the covariates that still play a role in $Q - Q_0$.

One could use this estimating function to define a closed form asymptotically efficient double robust estimator ψ_{DR} defined as the solution of the efficient influence curve estimating equation,

$$0 = P_n D^*(Q_n, g_n, \psi),$$

given estimators Q_n of Q_0 and g_n of g_0.

We can also compute a collaborative double robust asymptotically efficient targeted maximum likelihood estimator, which has various previously presented advantages: in particular, it is guaranteed to be a substitution estimator, and it will only pursue adjustment in g_n that remains helpful after the adjustment carried out by Q_n, thereby resulting in more effective adjustment sets and bias reduction.

A targeted maximum likelihood estimator is a substitution estimator $\Psi(\hat{P}^*)$, where the estimated data generating distribution $\hat{P}^*$ is such that it solves the efficient influence curve estimating equation,

$$0 = P_n D^*(Q(\hat{P}^*), g(\hat{P}^*), \Psi(\hat{P}^*)).$$

As a consequence, the substitution estimator (TMLE) $\Psi(\hat{P}^*)$ is double robust and efficient, and collaborative double robust if one uses the collaborative targeted maximum likelihood estimator that builds g_n based on the log-likelihood of Q_0, as presented in detail in van der Laan and Gruber (2010).

Targeted maximum likelihood estimator. Let's now explain the targeted maximum likelihood algorithm that maps an initial estimator $\hat{P}$ into a targeted fit $\hat{P}^*$. Suppose Y is binary. Given an initial estimator Q_n of $E(Y \mid A, W)$, an initial estimator g_n of $P(A \mid W)$, empirical distribution of W, in order to compute the targeted MLE, we define the fluctuation $\text{logit}Q_n(\epsilon_1)(A, W) = \text{logit}Q_n(A, W) + \epsilon_1 C_1(g_n)(A, W)$, and $\text{Logit}(g_n(\epsilon_2)(0 \mid W)) = \text{Logit}(g_n(0 \mid W)) + \epsilon_2 C_2(\hat{P})(W)$, where these two clever covariates are defined as

$$\begin{aligned} C_1(g) &= \left\{ \frac{I(A=1)}{P(A=0)} \frac{g(0 \mid W)}{g(1 \mid W)} - \frac{I(A=0)}{P(A=0)} \right\} \\ C_2(P) &= \frac{1}{P(A=0)} \left\{ Q(P)(1, W) - Q(P)(0, W) - \Psi(P) \right\}. \end{aligned}$$

These two one-dimensional fluctuations of the regression Q_n and the treatment mechanism g_n represents a fluctuation $\hat{P}(\epsilon)$ of $\hat{P}$, where the empirical distribution of W is hold fixed: the empirical distribution is already unbiased for the parameter of interest so that no fluctuation is needed. We estimate ϵ with maximum likelihood: note that ϵ_1 is estimated with standard linear logistic regression fixing Q_n as an offset, and ϵ_2 is estimated with standard linear logistic regression fixing $g_n(0 \mid W)$ as offset in the logistic regression model for $P(A = 0 \mid W)$.

This maximum likelihood estimator $\epsilon_n = (\epsilon_{1n}, \epsilon_{2n})$ now defines an update $\hat{P}^1 = \hat{P}(\epsilon_n)$. This targeted maximum likelihood updating is iterated till convergence and the final $\hat{P}^*$, identified by a Q_n^*, g_n^* (and the empirical for P_W), is called the targeted maximum likelihood estimator of the distribution P_0, while $\Psi(\hat{P}^*)$ is called the targeted maximum likelihood estimator of ψ_0. We have that the targeted maximum likelihood estimator $\Psi(\hat{P}^*)$ solves the efficient influence curve estimating equation, as presented above. We can use machine learning/super learning to obtain the initial $\hat{P}$ (i.e. Q_n and g_n).

Since $\hat{P}^*$ solves, in particular,

$$0 = \frac{1}{n} \sum_{i=1}^{n} \frac{I(A_i = 0)}{P(A=0)} \left\{ Q_n^*(1, W_i) - Q_n^*(0, W_i) - \Psi(\hat{P}^*) \right\},$$

it follows that the targeted MLE $\Psi(\hat{P}^*)$ can also be evaluated as

$$\Psi(\hat{P}^*) = \hat{E}(Q_n^*(1, W) - Q_n^*(0, W) \mid A = 0),$$

i.e., as the empirical mean of $Q_n^*(1, W) - Q_n^*(0, W)$ among the observations with $A_i = 0$. Apparently, in this evaluation g_n^* can be ignored.

Collaborative targeted MLE. The collaborative double robustness of the efficient influence curve allows us to also implement the collaborative targeted MLE of van der Laan and Gruber (2010). In this case, given the initial estimator Q_n, one starts with a g_n being an intercept model, and one selects the main term extensions of g_n that yields the maximal gain in log-likelihood of the Q-factor during the targeted maximum likelihood algorithm that starts at Q_n and this extension g_n. This process is iterated thereby building a main term regression model for g_0 that is based on the log-likelihood of Q. If there is no main term extension that improves the log-likelihood, then one carries out the previous TMLE update of Q_n using the previous g-fit, and one starts extending the current g-fit based on the log-likelihood of the TMLE-update of the just updated Q_n, so that the log-likelihood of Q always increases during these steps. This generates a sequence of targeted maximum likelihood estimators indexed by the number of moves that were used to build the g-fit. The number of moves used to build g is selected with likelihood based cross-validation, possibly penalizing the cross-validated log-likelihood as proposed in van der Laan and Gruber (2010).

Many variations of this collaborative TMLE algorithm can be considered. The main terms can include propensity score dimension reductions indexed by different adjustment sets, so that the above algorithm is still arbitrarily nonparametric. The common goal is to generate a sequence of targeted maximum likelihood estimators (Q_n^{j*}, g_n^{j*}) corresponding with starting estimators (Q_n^j, g_n^j), where the log-likelihood of Q_n^{j*} is increasing in j, and g_n^j (and thereby g_n^{j*}) is increasingly nonparametric. The choice for the number of moves j is then selected based on likelihood based cross-validation.

2.6 Designing the two sample study.

We now understand the estimator as an estimate of the statistical parameter ψ_0, and we also understand under what condition this statistical parameter ψ_0 equals the wished additive causal effect ψ_0^F. From that we conclude that it is important to measure all individual characteristics that can explain the effect of differences in the environments (i.e., e_0, e_1) of the two populations on the individual outcome, so that units in population 1 with $W = w$ are exchangeable with units in population 0 with $W = w$, w.r.t. the counterfactual outcome distributions. However, if e_0 is very different from e_1, then the covariate distributions P_{e_0} and P_{e_1} will be different, thereby possibly generating lack of experimentation for g_0: i..e $g_0(1 \mid W)$ gets close to 0 or 1 for some W-values. This increases the asymptotic variance of the targeted MLE, even if it does not result in non-identifiability: in other words, the variance of the efficient influence curve for ψ_0 increases when the covariate distributions P_{e_0} and P_{e_1} are getting more separated. As a consequence, even if all the wished W can be measured so that the effect of E on Y can be blocked, it is still very crucial

that the two populations are quite comparable w.r.t. to the factors e that have an impact on the outcome. The better job one does on that front, the smaller the asymptotic variance of the targeted MLE adjusting for W will be. This raises the issue of using a matched cohort design, involving matching a unit from the treated population with a unit from the control population based on a set of variables that are not affected by the treatment.

Two target parameters: average causal effect, and average causal effect among treated. Recall that we defined a random variable $O = (B, W = W(B), Y = Y(B)) \sim P_O$ representing the data on a random draw from the two populations combined, and representing the two sample problem of sampling n_0 observations from $P_{a_0=0,e_0}$ and n_1 observations from $P_{a_1=1,e_1}$. We work with the random experiment defined by O because it allows us to view the data set as one sample of i.i.d observations, while we fully respect the true two sample estimation problem. The model $\mathcal{M}$ for P_O is nonparametric.

We will be considering two target statistical parameters of P_O:

$$\begin{aligned}\psi_0 &= E_W\{E(Y \mid B = (1, e_1), W) - E(Y \mid B = (0, e_0), W)\} \\ \psi_0^t &= E(E(Y \mid B = (1, e_1), W) - E(Y \mid B = (0, e_0), W) \mid B = (1, e_1)).\end{aligned}$$

Theorem 1 proves that under an NPSEM in which (E, W, A) is randomized and in which Y is only affected by E through W, we have $\psi_0 = E\{Y(1) - Y(0)\}$ is the additive causal effect of A, one would obtain if one would be able to randomize A individually to each unit in the combined population and take a difference in means for the two samples, and sample size is infinity. Similarly, Theorem 2 shows that under these same assumptions $\psi_0^t = E(Y(1) - Y(0) \mid B = (1, e_1))$ is an additive causal effect of treatment for population 1, i.e. the treated population. As we will see the latter parameter is easier to identify from the data and makes the matched cohort design (defined below) particularly effective and optimal.

The efficient influence curves for ψ_0 and ψ_0^t are given by

$$\begin{aligned}D^*(Q_0, g_0, \psi_0)(O) &= \left\{\frac{I(B=1)}{g_0(1 \mid W)} - \frac{I(B=0)}{g_0(0 \mid W)}\right\}(Y - Q_0(B, W)) \\ &\quad + Q_0(1, W) - Q_0(0, W) - \Psi(Q_0) \\ D^{*t}(Q_0, g_0, \psi_0^t) &= \left\{\frac{I(B=1)}{P(B=1)} - \frac{I(B=0)}{P(B=1)}\frac{g_0(1 \mid W)}{g_0(0 \mid W)}\right\}(Y - Q(B, W)) \\ &\quad + \frac{I(B=1)}{P(B=1)}(Q(1, W) - Q(0, W) - \Psi(P_0)).\end{aligned}$$

Matched cohort sampling: Instead of the two sample design considered above which we treat as the equivalent of sampling $n_0 + n_1$ i.i.d. copies of $O = (B, W, Y)$, we can also consider the following matched cohort sampling:

- Let $M \subset W$ be a subset of the covariates W which represents the variable we will match on.
- Sample $W(e_1), Y(1, e_1)$ from the conditional distribution of (W, Y), given $B = (1, e_1)$. Let m_1 denote the observed value of M_1: i.e., $M_1 = m_1$.
- Sample J times $W(e_0), Y(0, e_0)$ from the conditional distribution (W, Y), given $B = (0, e_0)$ and $M = m_1$.
- Let
$$O^m \equiv \left((W(e_1), Y(1, e_1)), (W(e_0)^j, Y(0, e_0)^j), j = 1, \ldots, J\right),$$
be the cluster of matched observations.
- Repeat this experiment n times, resulting in n clusters O_i^m, $i = 1, \ldots, n$.
- It is noted that the dependence of observations within a cluster is only due to the matching on variable M: e.g., if the matching variable is empty, each cluster consists of i.i.d. copies.

2.7 Estimation in matched cohort designs.

Matched cohort designs provide a biased sample from the distribution of $O = (B, W, Y)$, so that a new identifiability result is required: we only provided the identifiability based on sampling i.i.d. copies of O. Targeted maximum likelihood estimation and efficient estimation in general, based on this type of case-control sampling, including matched case-control/cohort sampling, was studied in van der Laan (2008) and Rose and van der Laan (2008). In this work it is assumed that the following quantities are known

$$\begin{aligned} q_0 &= P(B = 1)) = \frac{n_1}{n_0 + n_1} \\ \bar{q}_0(M) &= \frac{q_0}{P(B = 1 \mid M)} P(B = 0 \mid M). \end{aligned}$$

The knowledge of these quantities allows one to identify any parameter that would have been identifiable under regular i.i.d sampling of $O = (B, W, Y)$. Therefore, this knowledge allows us to target the causal effect parameters ψ_0 and ψ_0^t of interest.

The case-control weighted targeted MLE is now defined by applying the targeted MLE of ψ_0 or ψ_0^t presented above, based on i.i.d. sampling of (B, W, Y), but giving each observation a weight. The observations with $B_i = 1$ are assigned the weight $q_0 = n_1/(n_0 + n_1)$. The J_i observations with $B_i = 0$ that were matched to a $B_i = 1$, receive weight $\bar{q}_0(M_i)/J$. The resulting case-control weighted targeted MLE now targets the same parameter ψ_0 or ψ_0^t, is asymptotically efficient, and it has the same double robustness property as the targeted MLE applied to the i.i.d. (B, W, Y). In addition, the efficient influence curve of ψ_0, ψ_0^t for this matched cohort sampling

model is given below, and can be used for statistical inference based on the case-control weighted targeted MLE as usual.

The kind of knowledge needed to determine these weights to correct for the matched sampling. Suppose one can determine for each matching category m, the proportion of units that have $M = m$ in the two populations/communities. This yields, $P(M = m|B = 1)$, $P(M = m|B = 0)$ for each m. In addition, we can set $P(B = 1) = n_1/n$, which corresponds with $P(E = e_1) = n_1/n$ in the NPSEM and thereby affects the interpretation of the marginal causal effects. This particular choice corresponds with the sampling actually used, and thereby is well supported by the data, but other choices can be accommodated as well. For example, if one aims to target the combined population, while n_1, n_0 are not proportional to population size, then $P(B = 1)$ is different from n_1/n. Off course, the required weights $q_0(M)$ are now determined by Bayes rule.

2.8 Evaluating gain of matching cohorts, relative to no-matching of the two cohorts.

In van der Laan (2008) it is shown that the efficient influence curve for the parameter ψ_0 based on sampling the cluster O^m equals a "case-control"-weighted efficient influence curve for the parameter ψ_0 based on sampling the data structure O. That is,

$$\begin{aligned} D^m(Q,g,\psi_0)(O^m) &= q_0 D^*(Q,g,\psi_0)(1,W(e_1),Y(1,e_1)) \\ &\quad + \frac{\bar{q}_0(M(e_1))}{J}\sum_{j=1}^{J} D^*(Q,g,\psi_0)(0,W(e_0^j),Y(0,e_0^j)) \\ D^{tm}(Q,g,\psi_0^t) &= q_0 D^{*t}(Q,g,\psi_0^t)(1,W(e_1),Y(1,e_1)) \\ &\quad + \frac{\bar{q}_0(M(e_1))}{J}\sum_{j=1}^{J} D^{*t}(Q,g,\psi_0^t)(0,W(e_0^j),Y(0,e_0^j)). \end{aligned}$$

This design includes the "no-matching" choice by setting M equal to empty set, and $J = n_0/n_1$, in which case $\bar{q}_0(M) = 1 - q_0$, and the case and control observations in the cluster are now independent. That is, if we set M empty, then this design corresponds with our original two sample study design we started out with.

Evaluating the benefit of matching in the design. The reader needs to recall that the variance of the efficient influence curve for a parameter (e.g) ψ_0 is the information bound for that parameter in the semiparametric model: as a consequence, any regular estimator has a larger asymptotic variance than the variance of the efficient influence curve, and an estimator is asymptotically efficient if and only if it is asymptotically linear with influence curve equal to the efficient influence curve.

Therefore, by studying the variance of the efficient influence curve of D^m, D^{tm} we can investigate if matching does decrease the variance relative to the no-matching design, and thereby increases the amount of information generated by the matching design for the purpose of estimation of ψ_0, ψ_0^t.

To consider the comparison of a matching design with no-matching, we focus on the case that $J = 1$, since the argument should not depend on the number of controls that are matched to the case. The efficient influence curve for ψ_0 based on sampling the cluster observation O^m is given by

$$D^m = q_0 \left\{ \frac{1}{g_0(1|W(e_1))}(Y(1,e_1) - Q_0(1,W(e_1))) + Q_0(1,W(e_1)) - Q_0(0,W(e_1)) - \Psi(Q_0) \right\}$$
$$+\bar{q}_0(M(e_1)) \left\{ -\frac{1}{g_0(0|W(e_0))}(Y(0,e_0) - Q_0(0,W(e_0))) + Q_0(1,W(e_0)) - Q_0(0,W(e_0)) - \Psi(Q_0) \right\}.$$

The efficient influence curve for ψ_0^t based on sampling the cluster observation O^m is given by:

$$-D^{tm} = q_0 \left\{ -\frac{1}{P(B=1)}(Y(1,e_1) - Q(1,W(e_1))) + \frac{1}{P(B=1)} \{Q(0,W(e_1)) - Q(1,W(e_1)) + \Psi(P)\} \right\}$$
$$+\bar{q}_0(M(e_1)) \left\{ \frac{1}{P(B=1)} \frac{g_0(1|W(e_0))}{g_0(0|W(e_0))}(Y(0,e_0) - Q(0,W(e_0))) \right\}.$$

Firstly, it is good to see that indeed, if M is empty, and thereby $\bar{q}_0(M) = 1 - q_0$, then D^m, D^{tm} correspond with i.i.d sampling of $O = (B, W, Y)$ and corresponding efficient influence curve D^*, D^{*t}, for ψ_0 and ψ_0^t given above: D^m, D^{mt} just combine observations from control and treatment sample in the cluster O^m, but since the observations in a cluster are now independent, this coupling serves no purpose beyond that it allows us to compare the efficient influence curve with matching with the efficient influence curve for the original no-matching two sample design. Therefore, the question "Is matching improving the design w.r.t. target parameter?" corresponds with "Is the variance of the efficient influence curve D^m, D^{tm} smaller when M is close to W, relative to M is low dimensional, with the extreme being that M is empty.

To answer this question we simply write down the efficient influence curves for M being empty and $M = W$ for both target parameters. For that purpose we denote $q_0 = g_0(1)$, $P(B = 1 \mid M) = g_0(1 \mid M)$ to stress that these are related to the conditional probability distribution $g_0(\cdot \mid W) = P_0(B = \cdot \mid W)$ of B, given W. This means that we can denote $\bar{q}_0(M) = \{g_0(1)/g_0(1 \mid M)\}g_0(0 \mid M)$.

Additive Causal effect, no matching:

$$\begin{aligned} D^m \;=\; & \frac{g_0(1)}{g_0(1 \mid W(e_1))}(Y(1,e_1) - Q_0(1,W(e_1))) \\ & +g_0(1)\{Q_0(1,W(e_1)) - Q_0(0,W(e_1)) - \Psi(Q_0)\} \\ & -\frac{1 - g_0(1)}{g_0(0 \mid W(e_0))}\{Y(0,e_0) - Q_0(0,W(e_0))\} \\ & +\frac{1 - g_0(1)}{g_0(0 \mid W(e_0))}\{Q_0(1,W(e_0)) - Q_0(0,W(e_0)) - \Psi(Q_0)\}. \end{aligned}$$

Causal effect, matching (M):

$$\begin{aligned} D^m &= \frac{g_0(1)}{g_0(1 \mid W(e_1))}(Y(1,e_1) - Q_0(1,W(e_1))) \\ &\quad + g_0(1)\left\{Q_0(1,W(e_1)) - Q_0(0,W(e_1)) - \Psi(Q_0)\right\} \\ &\quad - \frac{g_0(1)g_0(0 \mid M(e_1))}{g_0(1 \mid M(e_1))g_0(0 \mid W(e_0))}(Y(0,e_0) - Q_0(0,W(e_0))) \\ &\quad + \frac{g_0(1)g_0(0 \mid M(e_1))}{g_0(1 \mid M(e_1))}\left\{Q_0(1,W(e_0)) - Q_0(0,W(e_0)) - \Psi(Q_0)\right\}. \end{aligned}$$

Causal effect, full matching ($M = W$):

$$\begin{aligned} D^m &= \frac{g_0(1)}{g_0(1 \mid W(e_1))}(Y(1,e_1) - Q_0(1,W(e_1))) \\ &\quad + g_0(1)\left\{Q_0(1,W(e_1)) - Q_0(0,W(e_1)) - \Psi(Q_0)\right\} \\ &\quad - \frac{g_0(1)}{g_0(1 \mid W(e_0))}(Y(0,e_0) - Q_0(0,W(0,e_0))) \\ &\quad + g_0(0 \mid W(e_0))\left\{Q_0(1,W(e_0)) - Q_0(0,W(e_0)) - \Psi(Q_0)\right\}, \end{aligned}$$

where $W(e_0) = M(e_1) = W(e_1)$ with probability 1. We note that the inverse weighting by $g(0 \mid W)$ and $g(1 \mid W)$ is reduced to inverse weighting by $g(1|W)$ only, due the matching. Therefore, it seems that matching reduces the variance in many cases, and, at least, weakens the required identifiability condition to only $g_0(1|W) > 0$ a..e. Explicit calculations, not carried out here, will have to provide more support for this claim.

Causal effect among treated, No matching design:

$$\begin{aligned} -D^{tm} &= -(Y(1,e_1) - Q(1,W(e_1))) + \{Q(0,W(e_1)) - Q(1,W(e_1)) - \Psi(P)\} \\ &\quad + \frac{g_0(0)}{g_0(1)}\frac{g_0(1 \mid W(e_0))}{g_0(0 \mid W(e_0))}(Y(0,e_0) - Q(0,W(e_0))). \end{aligned}$$

Causal effect among treated, Matching design:

$$\begin{aligned} -D^{tm} &= -(Y(1,e_1) - Q(1,W(e_1))) + \{Q(0,W(e_1)) - Q(1,W(e_1)) + \Psi(P)\} \\ &\quad + \frac{g_0(0 \mid M(e_1))}{g_0(1 \mid M(e_1))}\frac{g_0(1 \mid W(e_0))}{g_0(0 \mid W(e_0))}(Y(0,e_0) - Q(0,W(e_0))). \end{aligned}$$

Causal effect among treated, full matching design ($M = W$):

$$\begin{aligned} -D^{tm} &= -(Y(1,e_1) - Q(1,W(e_1))) + \{Q(0,W(e_1)) - Q(1,W(e_1)) + \Psi(P)\} \\ &\quad + (Y(0,e_0) - Q(0,W(e_0))) \\ &= Y(1,e_1) - Y(0,e_0) - \Psi(P). \end{aligned}$$

Note that the estimator ψ_n^t that solves the efficient influence curve equation $P_n D^{tm}(Q_n, g_n, \psi^t) = 0$ is given by $\psi_n^t = P_n Y(1,e_1) - P_n Y(0,e_0)$, a difference of

sample means between the two groups, where the observation $Y(1, e_1), Y(0, e_0)$ are from a subject with the same covariate W. This suggests strongly that the efficient influence curve for the full-matching design has the smallest variance, thereby establishing the benefit of matching for the purpose of estimation of ψ_0^t.

Remark concerning matching in case-control sampling relative to matched cohort sampling. As shown in Rose and van der Laan (2009), by practical demonstration in simulation studies, case-control studies using matching carry typically less information about the parameter of interest than regular case-control designs. This can easily be seen by the approach followed above. Suppose the underlying data structure is (W, A, Y), $M \subset W$, and let ψ_0 be the target parameter with efficient influence curve D^* under i.i.d sampling of (W, A, Y), and nonparametric model for the distribution of (W, A, Y). Consider matched case-control sampling, conditioning on a binary variable R being a function of (W, A, Y) (playing the role of either Y or A say): one samples a "case", W, A, Y given $R = 1$, let $M_1 = m_1$, and subsequently one samples a control (W, A, Y), given $R = 0$ and $M = m_1$. By the general results in van der Laan (2008), the efficient influence curve under this matched-case-control sampling is given by $q_0 D^*(R = 1, W_1, A_1, Y_1) + \bar{q}_0(M_1) D^*(R = 0, W_0, A_0, Y_0)$. To make this matched sampling design effective one hopes to see that the weight $\bar{q}_0(M_1)$ cancels/stabilizes an inverse weight that appears in $D^*(R = 0, W_0, A_0, Y_0)$. Since the inverse weighting in D^* concerns inverse weighting by $P(A \mid W)$, multiplying by a $\bar{q}_0(M)$ that has a $P(A \mid M)$ in denominator can indeed do the job. Thus conditioning on $R = A$ and using matching can help. Indeed, above we determined that this happens in our setting with $R = A$. However, since the inverse weighting in D^* concerns inverse weighting by $P(A \mid W)$, using the matching when one conditions on Y, as in typical case-control studies, will never cancel or stabilize such weights, but induces additional unstable weighting by $\bar{q}_0(M)$ instead: Here one needs to note that in this case $\bar{q}_0(M) = P(Y = 1)P(Y = 0 \mid M)/P(Y = 1 \mid M)$ thus creating a singularity if $P(Y = 1 \mid M)$ can get small.

Summary regarding optimizing the design. If one can select two populations for which the pre-treatment covariate distributions P_{e_1}, P_{e_0} are almost equivalent, i.e. $P_{e_1}(W) \approx P_{e_0}(W)$, while W blocks any effect of E on Y, then that implies that $P(B = 1 \mid W) \approx P(B = 1)$, and thereby will result in an excellent information bound for any target parameter ψ_0. On the other hand, if this is not possible, then one still has the good option of using a matched cohort design, and targeting the causal effect for the treatment-population, ψ_0^t.

2.9 Extension to causal effect of treatment at time t.

We will now extend the above framework to causal effects of changes in treatment over time t.

Suppose we observe $n = n_1 + n_0$ i.i.d. copies of $(B, W(B), Y(B))$, where $B \in \{(e_0, a_0), (e_1, a_1)\}$, B is Bernoulli with probability n_1/n, conditional on $B = (e_1, a_1)$, $(W(B), Y(B))$ follows a distribution P_{e_1,a_1}, and, conditional on $B = (e_0, a_0)$, $(W(B), Y(B))$ follows a distribution P_{e_0,a_0}. Here W and Y are time-dependent processes over time $t = 1, \ldots, \tau$.

We assume an NPSEM:

$$\begin{aligned} U &\sim P_U \\ E &= f_E(U_E) \\ A &= f_A(E, M, U_A) \\ W(t) &= f_{W(t)}(M, \bar{Y}(t-1), \bar{W}(t-1), \bar{A}(t), \bar{E}(t), U_{W(t)}) \\ Y(t) &= f_{Y(t)}(M, \bar{Y}(t-1), \bar{W}(t), \bar{A}(t), \bar{E}(t), U_{Y(t)}) \\ t &= 1, \ldots, \tau. \end{aligned}$$

Here $E \in \{e_0, e_1\}$ and $A \in \{a_0, a_1\}$. The relation to the observed data is that the sampling distributions P_{e_0,a_0} and P_{e_1,a_1} are the distribution of the counterfactuals $(W(e_0, a_0), Y(e_0, a_0))$ and $(W(e_1, a_1), Y(e_1, a_1))$, respectively, defined by this NPSEM.

Let $Y^*(t)$ be an outcome of interest measured after $A(t)$. We define the following t-specific causal effects, $\psi_0(t)$ and $\psi_0^t(t)$ on the NPSEM:

$$\begin{aligned} \Psi^F(P_{U,X})(t) &\equiv EY^*_{\bar{A}(t-1)a_1(t)\bar{E}(t)}(t) - EY^*_{\bar{A}(t-1)a_0(t)\bar{E}(t)}(t) \\ \Psi^{F*}(P_{U,X})(t) &\equiv E\left\{Y^*_{\bar{A}(t-1)a_1(t)\bar{E}(t)}(t) - Y^*_{\bar{A}(t-1)a_0(t)\bar{E}(t)}(t) \mid (E, A) = (e_1, a_1)\right\}. \end{aligned}$$

This corresponds with only intervening on $A(t)$ by setting it at $a_1(t)$ and $a_0(t)$, resulting in two counterfactuals $Y(1)(t) = Y^*_{\bar{A}(t-1)a_1(t)\bar{E}(t)}(t)$, and $Y(0)(t) = Y^*_{\bar{A}(t-1)a_0(t)\bar{E}(t)}(t)$, so that these target parameters can also be denoted with $E(Y(1)(t) - Y(0)(t))$ and $E(Y(1)(t) - Y(0)(t) \mid (A, E) = (a_1, e_1))$.

Define

$$\begin{aligned} E^*(t) &\equiv \bar{A}(t-1), \bar{E}(t) \\ W^*(t) &\equiv \bar{W}(t-1), \bar{Y}(t-1). \end{aligned}$$

In addition, we have $A(t) \in \{a_0(t), a_1(t)\}$, and outcome $Y^*(t)$. The original NPSEM above, implies that these four variables also satisfy an NPSEM, in which one first

draws E^*, then $W^*(t)$, then $A(t)$, and finally $Y^*(t)$:

$$\begin{aligned} U &\sim P_U \\ E^*(t) &= f_{E^*(t)}(U_{E^*(t)}) \\ W^*(t) &= f_{W^*(t)}(E^*(t), U_{W^*(t)}) \\ A(t) &= f_{A(t)}(E^*(t), W^*(t), U_{A(t)}) \\ Y^*(t) &= f_{Y^*(t)}(E^*(t), W^*(t), A(t), U_{Y^*(t)}). \end{aligned}$$

This NPSEM for the endogenous nodes $(E^*(t), W^*(t), A(t), Y^*(t))$ defines counterfactual random variables $Y^*(A(t) = a_1(t))(t)$ and $Y^*(A(t) = a_0(t))(t)$, by intervening on $A(t)$, and counterfactual random variables $Y^*(e^*(t), a(t))(t)$, by intervening on $e^*(t), a(t)$.

Our observed data corresponds with observing $n = n_1+n_0$ i.i.d. $(B, W^*(B)(t), Y^*(B)(t))$, where $B \in \{(e_0, a_0), (e_1, a_1)\}$ and $P(B = (e_1, a_1)) = P(E^*(t) = e_1^*(t))$ with $e_1^*(t)$ deterministically determined by $(E, A) = (e_1, a_1)$, and, conditional on $B = (e_0, a_0)$, $(W^*(B)(t), Y^*(B)(t))$ is distributed as its counterfactual analogue corresponding with setting $(E, A) = (e_0, a_0)$, and thereby setting $E^*(t) = e_0^*(t)$ and $A(t) = a_0(t)$. Similarly, the above statement applies conditional on $B = (e_1, a_1)$.

We have now reformulated the causal effect estimation problem as a t-specific version of the problem addressed in previous subsections. As a consequence, we can apply Theorem 1 to this t-specific identification problem, which proves that, if $(E^*(t), W^*(t), A(t))$ is independent of the counterfactuals $Y^*(e^*(t), w^*(t), a(t))(t)$, and $Y^*(t) = f_{Y^*(t)}(W^*(t), A(t), U_{Y^*(t)})$ is not a function of $E^*(t)$, then

$$\begin{aligned} \Psi(P_{U,X})(t) &= E(Y(1)(t) - Y(0)(t)) \\ &= E_{W^*(B)(t)} E(Y^*(B)(t) \mid B = (e_1, a_1), W^*(B)(t)) \\ &\quad - E_{W^*(B)(t)} E(Y^*(B)(t) \mid B = (e_0, a_0), W^*(B)(t)) \\ &\equiv \Psi(P_0)(t) \end{aligned}$$

This statistical parameter of the distribution P_0 of the data $O = (B, W^*(B)(t), Y^*(B)(t))$ can be double robust and efficiently estimated with targeted MLE as in previous subsections.

One might also be concerned with estimation of a weighted average across time t of $\Psi(P_0)(t)$. In that case, one can substitute the targeted MLE for $\Psi(P_0)(t)$ for each t, or, one can apply a single targeted MLE targeting the weighted average to a pooled sample that includes τ t-specific records $(t, W^*(t), Y^*(t))$ for each subject.

Similarly, we have this result for the t-specific causal effect among the treated population,

$$\begin{aligned} \Psi^*(P_{U,X})(t) &= E(Y(1)(t) - Y(0)(t) \mid (E, A) = (e_1, a_1)) \\ &= E\left\{E(Y^*(B)(t) \mid B = (e_1, a_1), W^*(B)(t)) \mid B = (e_1, a_1)\right\} \\ &\quad - E\left\{E(Y^*(B)(t) \mid B = (e_0, a_0), W^*(B)(t)) \mid B = (e_1, a_1)\right\} \\ &\equiv \Psi^*(P_0)(t) \end{aligned}$$

Again, we already developed the targeted MLE for this target parameter of P_0 in the previous subsections, and the same remark as above applies.

For $\psi_0^*(t)$ or $\psi_0(t)$ we could employ the matched sampling design, matching on a subset of $W^*(t)$. However, that design would then be targeted towards this particular t-specific effect. If one is concerned with estimation of an average of $\psi_0^*(t)$, then, one should only match on covariates that are not affected by any of the $A(t)$: i.e., one would then match on the covariates M that were realized before one assigned the treatment regimen A, or, more general, are not affected by the treatment regimen.

Again, the combination of using matching in the design and the adjustment by $W^*(t)$ carried out by targeted MLE provides a good way to deal with the different environmental factors e_0 and e_1 in the two samples.

3 Assigning two interventions to multiple populations.

The above framework can be generalized to handle multiple populations. The theorem below, which generalizes Theorem 1, proves that nothing fundamental changes: the well defined causal effects on the NPSEM allow an identifiability result under the same assumptions as stated, and subsequently one applies targeted MLE to estimate these statistical parameters. To achieve these assumptions one wants to collect individual covariates that can block effect of the environmental factors that are different between populations, and for the sake of optimizing the design, one also wants to match on such individual covariates across the multiple populations.

Theorem 3 NPSEM. *Consider a NPSEM with structural equations for the endogenous $X = (E, W, A, Y)$,*

$$\begin{aligned} E &= f_E(U_E) \\ W &= f_W(E, U_W) \\ A &= f_A(E, W, U_A) \\ Y &= f_Y(E, W, A, U_Y), \end{aligned}$$

and exogenous $U = (U_E, U_W, U_A, U_Y)$. Let $A \in \{0,1\}$, $E \in \{e_1, \ldots, e_J\}$. Let α be the marginal probability distribution of E.

Counterfactuals. *Let $Y(1) = f_Y(E, W, 1, U_Y)$ and $Y(0) = f_Y(E, W, 0, U_Y)$ denote the counterfactuals corresponding with setting $A = 1$ and $A = 0$, respectively. We also define $(W(e_j), Y(1, e_j))$ and $(W(e_j), Y(0, e_j))$ as the post-intervention random variable corresponding with setting $A = 1, E = e_j$ and $A = 0, E = e_j$, respectively, $j = 1, \ldots, J$. We also define $Y(e, w, a) = f_Y(e, w, a, U_Y)$ as the post-intervention counterfactual of Y corresponding with intervention $E = e, W = w, A = a$, $e \in \{e_1, \ldots, e_J\}$, $a \in \{0,1\}$, and all possible w.*

Observed multi-sample data set. *Let $\mathcal{B} = ((a_j, e_j) : j = 1, \ldots, J)$ be a set with J given treatment and exposure combinations. We denote the distributions of*

the corresponding counterfactuals $(W(e_j), Y(a_j, e_j))$ *with* P_{a_j,e_j}, $j = 1, \ldots, J$. *We observe* n_j *i.i.d. observations from* P_{a_j,e_j}, $j = 1, \ldots, J$.

Reformulation of observed data. *Let* $O = (B, W(B) \equiv W(e_B), Y(B))$, *where* $B \sim \alpha$, $B \in \mathcal{B}$, *conditional on* $B = (a_j, e_j)$, O *is distributed as* $(W(e_j), Y(a_j, e_j)) \sim P_{a_j,e_j}$, $j = 1, \ldots, J$. *Let* $P(B = (a_j, e_j)) = \alpha(j)$, $j = 1, \ldots, J$, *be the marginal proability distribution of* B. *We note that* $P(W(B) = w) = \sum_{j=1}^{J} P_{W(e_j)}(w)\alpha(j)$. *Let* P_O *be the probability distribution of* O: $P_O = P_O(P_{U,X})$.

Relevance to multi-sample data set. *We note that the distribution of* P_O *also approximates the multi-sample data structure in which one samples* n_j *i.i.d. observations from* P_{a_j,e_j}, $j = 1, \ldots, J$, *by setting* $\alpha(j) = n_j / \sum_j n_j$.

Target parameter on NPSEM. *Consider the following parameter of the distribution of* (U, X):

$$\Psi^F(P_{U,X}) = EY(1) - EY(0),$$

the additive causal effect of setting $A = 1$ *versus* $A = 0$.

Exclusion and Randomization assumption on NPSEM. *Assume that* Y *is only a function of* E *through* W, *i.e.,* $Y = f_Y(W, A, U_Y)$ *in the NPSEM, and that the distribution of* $U = (U_E, U_W, U_A, U_Y)$ *is such that* (E, W, A) *is independent of* $Y(e, w, a)$ *for all* e, w, a.

Identifiability result. *Let* $\mathcal{B}_0, \mathcal{B}_1$ *be the partitioning of* $\mathcal{B}$ *into the regimens with* $a_j = 0$ *and the regimens with* $a_j = 1$, *respectively. Then, for any* $(a_j, e_j) \in \mathcal{B}_1$, *we have*

$$EY(1) = E_{W(B)}E(Y(B) \mid W(B), B = (a_j, e_j)),$$

and, for any $(a_j, e_j) \in \mathcal{B}_0$, *we have*

$$EY(0) = E_{W(B)}E(Y(B) \mid W(B), B = (a_j, e_j)).$$

In particular,

$$\begin{aligned} EY(1) - Y(0) &= \frac{1}{\mid \mathcal{B}_1 \mid} \sum_{b \in \mathcal{B}_1} E_{W(B))}E(Y(B) \mid W(B), B = b) \\ &\quad - \frac{1}{\mid \mathcal{B}_0 \mid} \sum_{b \in \mathcal{B}_0} E_{W(B)}E(Y(B) \mid W(B), B = b). \end{aligned}$$

Most importantly, we have the following identifiability result:

$$\begin{aligned} E(Y(1) \mid W = w) &= E(Y(B) \mid B \in \mathcal{B}_1, W(B) = w) \\ E(Y(0) \mid W = w) &= E(Y(B) \mid B \in \mathcal{B}_0, W(B) = w), \end{aligned}$$

and thereby

$$E(Y(1) - Y(0)) = E_{W(B)}E(Y(B) \mid B \in \mathcal{B}_1, W(B)) - E(Y(B) \mid B \in \mathcal{B}_0, W(B)). \quad (3)$$

Proof. Firstly, for the full data parameter, we have

$$\begin{array}{l}\Psi_1^F(P_{U,X}) = EY(1) = Ef_Y(1, W, U_Y) \\ = \sum_w E(f_Y(1, w, U_Y) \mid W = w)P(W = w) \\ = \sum_w Ef_Y(1, w, U_Y)P(W = w),\end{array}$$

where we used at the last equality that W is independent of $Y(e, w, a)$, by assumption. We note that $P(W = w) = \sum_{j=1}^J P(W(e_j) = w \mid E = e_j)P(E = e_j) = \sum_{j=1}^J P_{W(e_j)}(w)\alpha_E(j)$.

Consider now the parameter $\Psi_j(P_0) = E_{W(B)}E(Y(B) \mid W(B), B = (1, e_j))$ of the distribution of observed data structure O. Since, given $B = (1, e_j)$, $(W(B), Y(B))$ is distributed as $(W(e_j), Y(1, e_j))$, we have

$$\begin{array}{rcl}E(Y(B) \mid W(B) = w, B = (1, e_j)) & = & E(f_Y(1, w, U_Y) \mid W(e_j) = w) \\ & = & E(f_Y(1, w, U_Y) \mid A = 1, E = e_j, W = w) \\ & = & Ef_Y(1, w, U_Y),\end{array}$$

where the second equality is implied by (A, E) being independent of $Y(e, w, a)$, given W, and the third equality is implied by (E, W, A) being independent of $Y(e, w, a)$, both consequences of our strong randomization assumption.

In addition, the observed data parameter $\Psi_j(P_0)$ for $EY(1)$ involves averaging w.r.t $P(W(B) = w) = \sum_{j=1}^J P_{W(e_j)}(w)P(B = (a_j, e_j))$. By assumption, $P(B = (a_j, e_j)) = P(E = e_j) = \alpha_E(j)$.

Thus, we conclude

$$\Psi_j(P_0) = \sum_w \{Ef_Y(A = 1, w, U_Y) - Ef_Y(A = 0, w, U_Y)\}P(W = w) = \Psi_1^F(P_{U,X}).$$

This completes the proof that the full data parameter Ψ_1^F of the distribution of (U, X), as defined by the NPSEM, can be identified as a mapping Ψ applied to the observed data distribution O implied by the distribution of (U, X).

We will now prove the last statement. We have

$$
\begin{aligned}
&E(Y(B) \mid B \in \mathcal{B}_1, W(B) = w) = \textstyle\sum_y y P(Y(B) = y \mid B \in \mathcal{B}_1, W(B) = w) \\
&= \textstyle\sum_y y \frac{P(Y(B)=y, B \in \mathcal{B}_1, W(B)=w)}{P(B \in \mathcal{B}_1, W(B)=w)} \\
&= \textstyle\sum_y y \frac{\sum_{b \in \mathcal{B}_1} P(Y(b)=y, B=b, W(b)=w)}{\sum_{b \in \mathcal{B}_1} P(W(b)=w, B=b)} \\
&= \textstyle\sum_y y \frac{\sum_{b \in \mathcal{B}_1} P(Y(b)=y \mid B=b, W(b)=w) P(W(b)=w, B=b)}{\sum_{b \in \mathcal{B}_1} P(W(b)=w, B=b)} \\
&= \textstyle\sum_y y \frac{\sum_{b=(1,e) \in \mathcal{B}_1} P(Y(1,e)=y \mid W(e)=w) P(W(b)=w, B=b)}{\sum_{b \in \mathcal{B}_1} P(W(b)=w, B=b)} \\
&= \textstyle\sum_y y \frac{\sum_{b=(1,e) \in \mathcal{B}_1} P(Y(1,w)=y \mid W(e)=w) P(W(b)=w, B=b)}{\sum_{b \in \mathcal{B}_1} P(W(b)=w, B=b)} \\
&\text{by exclusion restriction assumption} \\
&= \textstyle\sum_y y \frac{\sum_{b=(1,e) \in \mathcal{B}_1} P(Y(1,w)=y) P(W(b)=w, B=b)}{\sum_{b \in \mathcal{B}_1} P(W(b)=w, B=b)} \\
&\text{by strong randomization assumption} \\
&= \textstyle\sum_y y P(Y(1,w) = y) \\
&= \textstyle\sum_y y (P(Y(1) = y \mid W = w) \\
&= E(Y(1) \mid W = w).
\end{aligned}
$$

This completes the proof. □

The analogue theorem for the causal effect among the treated is generalized in the same way.

3.1 Efficient influence curve, Targeted ML Estimation, and statistical inference

The last identifiability result stated in the theorem teaches us that, under the exclusion restriction and randomization assumption, the causal parameter $EY(a)$ corresponds with statistical parameter

$$EY(a) = \Psi_a(P_0) = E_{W(B)} E(Y(B) \mid B \in \mathcal{B}_a, W(B)).$$

Thus, we can identify the additive causal effect of the community based intervention, $EY(1) - Y(0)$, with the statistical target parameter

$$\Psi(P_0) = E_{W(B)} \{E(Y(B) \mid B \in \mathcal{B}_1, W(B)) - E(Y(B) \mid B \in \mathcal{B}_0, W(B))\}.$$

The efficient influence curve, targeted MLE, collaborative targeted MLE, and statistical inference based on an estimate of the efficient influence curve, has been presented earlier, and corresponds exactly with the statistical target parameter $E_W\{E(Y \mid A = 1, W) - E(Y \mid A = 0, W)\}$ based on observing n i.i.d. copies of (W, A, Y) with $Y = Y(B)$, $W = W(B)$, and $A = I(B \in \mathcal{B}_1)$. Thus, the practical conclusion is that one can create one combined sample from the J community-specific samples, reduce each observation to W, A, Y by ignoring the data on environmental factors, and apply the targeted MLE.

4 Estimation and inference without the exclusion restriction assumption

Suppose the exclusion restriction assumption fails to hold. Our proposed parameter of the distribution of O involves a difference of $E_{W(B)}E(Y(B) \mid B \in \mathcal{B}_a, W(B))$ for the treated $a = 1$ and control $a = 0$. This suggest that if A is independent of E (i..e., A is randomly assigned to communities), then the bias of this parameter due to violation of the exclusion restriction assumption will be decreasing in the number of communities J. In this section we aim to incorporate this residual confounding by differences in environments of the treated and untreated communities in the statistical inference. Our proposed targeted ML estimator is unchanged, with the only remark that, if the number of communities is quite large, we recommend including environmental factors in the definition of $W(B)$, so that they are also potentially used in the adjustment: the collaborative targeted MLE could be used to make this decision data adaptively. This section is thereby only concerned with understanding the target of this targeted MLE as a causal parameter, and taking into account its bias w.r.t. a wished causal effect by appropriately enlarging the variance estimator.

4.1 Testing the exclusion restriction assumption.

Consider two communities (a_1, e_1) and (a_2, e_2) which received the same treatment, so that $a_1 = a_2$. Under the exclusion restriction assumption we have that $E_{W(B)}E(Y \mid W(B), B = (a_1, e_1)) - E_{W(B)}E(Y \mid W(B), B = (a_2, e_2)) = 0$ for any such pair of communities with $a_1 = a_2$. We could estimate the two parameters using stratification by community when estimating $E(Y \mid W(b), B = b)$, constructing a t-statistic, and carry out a test of the null hypothesis that the difference equal zero. In particular, we could target the difference between the average of all $\Psi_{e_j,a_j}(P_0)$ across $\{j : a_j = 1\}$ and the average of all $\Psi_{e_j,a_j}(P_0)$ across $\{j : a_j = 0\}$, and carry out a single targeted maximum likelihood based test for testing that this difference equals zero, where the estimation of $E(Y \mid W(b), B = b)$ is stratified by the a-component.

We conclude that the exclusion restriction assumption is a testable assumption.

4.2 The wished causal target of TMLE without the exclusion restriction assumption, and estimation of standard error of TMLE relative to this causal target

The following theorem establishes the bias of the target parameter $\Psi(P_0)$, when not assuming the exclusion restriction assumption, w.r.t. a well defined causal effect of treatment, as a function of the number J of sampled communities and the degree of violation of the exclusion restriction assumption. In particular, it provides us with an augmentation of the variance of the previously presented targeted ML estimator that takes into account that we only observe a finite sample of J communities while the exclusion restriction assumption might be violated.

Theorem 4 NPSEM-J. *Consider a J-specicic NPSEM with structural equations for the endogenous $X^J = (E^J, W^J, A^J, Y^J)$,*

$$\begin{aligned} E^J &= f_{E^J}(U_{E^J}) \\ W^J &= f_W(E^J, U_W) \\ A^J &= f_A(E^J, W^J, U_A) \\ Y^J &= f_Y(E^J, W^J, A^J, U_Y), \end{aligned}$$

and exogenous $U^J = (U_{E^J}, U_W, U_A, U_Y)$. Let $A^J \in \{0,1\}$, $E^J \in \{e_1, \ldots, e_J\}$. Let α^J be the marginal probability distribution of E. We note that all random variables are indexed by J because we are concerned in behavior of target parameter and estimator for J large. We also note that the deterministic functions in the NPSEM for W, A, Y are the same for each J, and the corresponding exogenous errors (U_W, U_A, U_Y) also have a common distribution.

Counterfactuals. *Let $Y^J(1) = f_Y(E^J, W^J, 1, U_{Y^J})$ and $Y^J(0) = f_Y(E^J, W^J, 0, U_{Y^J})$ denote the counterfactuals corresponding with setting $A^J = 1$ and $A^J = 0$, respectively. We define $(W^J(e_j), Y^J(1, e_j))$ and $(W^J(e_j), Y^J(0, e_j))$ as the post-intervention random variable corresponding with setting $A^J = 1, E^J = e_j$ and $A^J = 0, E^J = e_j$, respectively, $j = 1, \ldots, J$. We also define $Y(e, w, a) = f_Y(e, w, a, U_Y)$ as the post-intervention counterfactual of Y^J (same for all J) corresponding with intervention $E^J = e, W^J = w, A^J = a$, $e \in \{e_1, \ldots, e_J\}$, $a \in \{0,1\}$, and all possible w.*

Observed multi-sample data set. *Let $\mathcal{B}^J = ((a_j, e_j) : j = 1, \ldots, J)$ be a set with J given treatment and exposure combinations. We denote the distributions of the corresponding counterfactuals $(W^J(e_j), Y^J(a_j, e_j))$ with $P^J_{a_j,e_j}$, $j = 1, \ldots, J$. We observe n_j i.i.d. observations from $P^J_{a_j,e_j}$, $j = 1, \ldots, J$. Let $n = \sum_{j=1}^J n_j$.*

Reformulation of observed data. *Suppose that $O^J = (B^J, W^J(B^J) \equiv W^J(e_{B^J}), Y^J(B^J))$ where $B^J \sim \alpha^J$, $B^J \in \mathcal{B}^J$, and conditional on $B^J = (a_j, e_j)$, O^J is distributed as $(W^J(e_j), Y^J(a_j, e_j)) \sim P^J_{a_j,e_j}$, $j = 1, \ldots, J$. Let $P(B^J = (a_j, e_j)) = \alpha^J(j)$, $j = 1, \ldots, J$, be the marginal probability distribution of B. We set $\alpha^J(j) = n_j/n$. We note that $P(W^J(B^J) = w) = \sum_{j=1}^J P_{W(e_j)}(w)\alpha^J(j)$, which equals $P(W^J = w)$ of the NPSEM. Let P_O be the probability distribution of O, which is identified by $P_{U,X}$: $P_O = P_O(P_{U,X})$.*

Relevance to multi-sample data set. *We note that the distribution of P_O also approximates the multi-sample data structure in which one samples n_j i.i.d. observations from P_{a_j,e_j}, $j = 1, \ldots, J$, by setting $\alpha^J(j) = n_j / \sum_j n_j$.*

Randomization assumptions. *Assume the distribution of $U^J = (U^J_E, U_W, U_A, U_Y)$ is such that (E^J, W^J, A^J) is independent of $Y(e, w, a)$ for all e, w, a. Assume that the realized $\{b_j : j = 1, \ldots, J\}$ are the outcomes of J times drawing from a distribution of (E, A) with E varying over a possibly infinite set and $A \in \{0,1\}$: at least, approximately for J large. Assume also that A is independent of E (e.g., in a randomized community intervention trial). As a consequence, the marginal distribution*

of B in our observed data formulation is distributed as the distribution of (E, A) with E independent of A.

Results. *Suppressing J, we have*

$$\begin{aligned} & E(Y(B) \mid B \in \mathcal{B}_a, W(B) = w) \\ &= \sum_y y \frac{\sum_{b \in \mathcal{B}_a} P(Y(b)=y|W(b)=w)P(W(b)=w,B=b)}{\sum_{b \in \mathcal{B}_a} P(W(b)=w,B=b)} \\ &= \frac{\sum_{b \in \mathcal{B}_a} E(Y(b)|W(b)=w)P(W(e)=w)\alpha(e)}{\sum_{b \in \mathcal{B}_a} P(W(e)=w)\alpha(e)}. \end{aligned}$$

Thus,

$$E_{W(B)}E(Y(B) \mid B \in \mathcal{B}_a, W(B)) \quad = \quad \frac{P^J_{E|A=a} \sum_w \bar{P}^J(w)\bar{Q}_{E,a}(w)P_E(w)\alpha(E)}{P^J_{E|A=a}P_E(w)\alpha(E)},$$

where $\bar{P}^J(w) = P(W(B) = w)$, $P_e(w) = P(W(e) = w)$, $\bar{Q}_b(w) = E(Y(B) \mid B = b, W(B) = w)$, and $P^J_{E|A=a}$ is the empirical distribution of the conditional distribution of E, given $A = a$, based on $B_j = (E_j, A_j)$, $j = 1, \ldots, J$.

We view this parameter as a function $\Phi_a(P^J_{E|A=a})$, treating it as random through the empirical distribution of E, given $A = a$, based on $B_j = (E_j, A_j)$, $j = 1, \ldots, J$. The statistical parameter

$$\Psi^J(P^J_0) \equiv E_{W(B)}E(Y(B) \mid B \in \mathcal{B}_1, W(B)) - E_{W(B)}E(Y(B) \mid B \in \mathcal{B}_0, W(B)) \quad (4)$$

equals

$$\bar{\Phi}(P^J_B) \equiv \Phi_1(P^J_{E|A=1}) - \Phi_0(P^J_{E|A=0}).$$

If the exclusion restriction assumption holds for the J-specific NPSEM, then this equals $EY(1) - Y(0)$. However, without this assumption, note that a difference of $\bar{\Phi}(P^J_B)$ from zero can be due to both a true treatment effect or a difference between $P^J_{E|A=1}$ and $P^J_{E|A=0}$.

Wished target parameter with no residual confounding due to environmental factors. *We define a wished target as its limit if $P^J_{E|A=a}$ is replaced by its limit $P_{E|A=a}$ for both $a \in \{0, 1\}$, while we keep $\bar{P}^J(w)$ fixed at J:*

$$\begin{aligned} & \bar{\Phi}(P_B) = \Phi_1(P_{E|A=1}) - \Phi_0(P_{E|A=1}) \\ &= \frac{P_{E|A=1} \sum_w \bar{P}^J(w)\bar{Q}_{E,1}(w)P_E(w)\alpha(E)}{P_{E|A=1}P_E(w)\alpha(E)} - \frac{P_{E|A=1} \sum_w \bar{P}^J(w)\bar{Q}_{E,0}(w)P_E(w)\alpha(E)}{P_{E|A=1}P_E(w)\alpha(E)}, \end{aligned}$$

where, by assumption, $P_{E|A=1} = P_E$.

Asymptotic (in J) linearity of statistical target parameter as estimate of wished target parameter: *We have*

$$\bar{\Phi}(P^J_B) - \bar{\Phi}(P_B) = \frac{1}{J} \sum_{j=1}^{J} IC(B_j) + o_P\left(\frac{1}{\sqrt{J}}\right),$$

where

$$\begin{aligned} IC(B_j) &= IC_1(B_j) - IC_0(B_j), \\ IC_a(B_j) &= I(A_j = a) \sum_w \left\{ \frac{f_{B_j}(w)}{P_B g_B(w)} - \frac{P_B f_B(w)}{P_B^2 g_B(w)} g_{B_j}(w) \right\}, \\ f_B(w) &= \bar{P}^J(w) \bar{Q}_B(w) P_E(w) \alpha(E), \\ g_B(w) &= P_E(w) \alpha(E). \end{aligned}$$

Thus, we have that $\bar{\Phi}(P_B^J) - \bar{\Phi}(P_B)$ *is approximately normally distributed with mean 0 and variance*

$$\sigma^2(J) = \frac{VARIC(B)}{J}. \tag{5}$$

In particular, if in the J-specific NPSEM, $Y = f_Y(W, A, U_Y)$ *does not depend on E, i.e., if the exclusion restriction holds, then* $\sigma^2(J) = 0$.

The variance of $\bar{\Phi}(P_B^J)$ *can be estimated as*

$$\frac{1}{J^2} \sum_{j=1}^{J} \hat{IC}(B_j)^2,$$

where $\hat{IC}$ *is obtained by replacing* P_B *by its empirical distribution of* $B_1, \ldots, B_J$*, and the functions* f_B *and* g_B *are estimated as well with their empirical counterparts.*

4.3 Variance estimation incorporating residual confounding due to violation of exclusion restriction assumption.

Consider the statistical parameter $\Psi^J(P_0^J) = \Psi^J(Q_0)$ (4). This statistical parameter and the estimator is not affected by the exclusion restriction assumption to be true or not. In the previous section we proposed a targeted MLE of this target parameter $\Psi^J(Q_0)$, which is a substitution estimator obtained by plugging in a $\bar{Q}_n^*$ estimator of $\bar{Q}_0(w, a) = E(Y(B) \mid W(B) = w, B \in \mathcal{B}_a)$ and the empirical distribution to estimate the distribution $W^J(B^J)$. Under regularity conditions, this targeted MLE $\Psi^J(Q_n^*)$ is an asymptotically linear estimator of ψ_0^J with an influence curve $IC^J(O^J)$, conditional on the values (e_j, a_j), $j = 1, \ldots, J$. Let $\hat{\sigma}^2 = 1/N \sum_{j,i} \{\hat{IC}^J(O_{ji}^J)\}^2$ be the estimated variance of the influence curve, so that $\hat{\sigma}^2/N$ is an estimate of the variance of $\Psi^J(Q_n^*)$.

Our wished target is $\psi_0 = \bar{\Phi}(P_B)$ as defined in the theorem. Note that $\Psi^J(Q_0) - \psi_0$ is random through $B_j = (E_j, A_j)$, $j = 1, \ldots, J$. This implies that $\Psi^J(Q_n^*) - \Psi^J(Q_0)$ and $\Psi^J(Q_0) - \psi_0$ are asymptotically uncorrelated, and thus independent (both are asymptotically normally distributed).

Therefore, we can conclude that

$$\Psi^J(Q_n^*) - \psi_0 \sim N\left(0, \frac{\hat{\sigma}^2}{N} + \sigma^2(J)\right) \text{ for } N \text{ and } J \text{ large,}$$

where $\sigma^2(J)$ is defined by (5), and it equals the contribution to the variance due to dependence of $\bar{Q}_0(E, W, A)$ on E (i.e., due to violation of the exclusion restriction assumption). Note that the variance-term $\sigma^2(J)$ is of the order $1/J$, while the variance of the targeted MLE is of the order N.

We conclude that, without assuming the exclusion restriction assumption, the above results can be used to not only target the parameter $\Psi^J(Q_0)$ (which equals $EY(1) - Y(0)$ under the exclusion restriction assumption), but target the target parameter ψ_0 that takes out the residual environmental confounding due to differences in the environmental factors between treated and non-treated communities that could not be explained by the individually measured covariates. Our proposed variance estimate naturally adapts in the sense that it approximates the variance of the targeted MLE of $\Psi^J(Q_0)$ if the exclusion restriction assumption holds, while it gets appropriately augmented otherwise.

5 Causal effect of changes in treatment over time when a single time-dependent treatment regimen is assigned to a population

Our previous results have applications that are easily overseen. For that purpose, we start this section with providing a general way to think of our templates and theorems. One observes a sample of observations from a probability distribution that is indexed by a choice of intervention and external environmental factors, across a collection of combined interventions and environmental values. These different probability distributions generate the same data structure, such as a vector (W, Y) of covariates and an outcome. The samples might be independent as in sampling from different populations, but, as we point out in this section, they can as well be dependent as in sampling the same group across time. The key is that one assumes that each of these probability distributions are treatment-environment-specific counterfactual distributions defined by intervening on a single NPSEM. We define treatment-specific counterfactual outcome distributions on this same NPSEM, which are defined as the outcome of the experiment that first draws randomly from the set of possible environments, subsequently draws a covariate from the environment specific distribution, sets treatment, and finally draws the outcome, given the environment, covariate, and treatment. Differences between these treatment-specific counterfactual outcome distributions define now interesting causal effects that are free from environmental confounding. Our identifiability results provide now the conditions under which these distributions and thereby the corresponding causal contrasts are identified from the data generating distributions. Specifically, we define statistical parameters that equal the wished causal effects under well understood conditions on the NPSEM and the sampling (e.g., number) of the intervention and environment values.

The different environments can correspond with different populations, different neighborhoods or communities, but also different time-points at which the sample was taken. The exclusion restriction assumption states that the effect of different environmental factors on the outcome only occurs through the individually measured covariates. For example, if one sampled from a population at two different time points (e.g., a year apart), either involving independent sampling or sampling of one cohort over two time points, then changes in outcome distribution over time, in the absence of a change of treatment, need to be completely explained by a change in covariate distribution.

5.1 A two time-point example.

Consider a study in which we observe a sample of subjects from a population and expose them in the first year to a treatment $A(1) = 0$ and the second year to a treatment $A(2) = 1$. The data on these n subjects can be coded as $(W_i(t), Y_i(t))$, $t \in \{1, 2\}$, $i = 1, \ldots, n$, where $W_i(t)$ denotes the individual history before $A(t)$, and $Y_i(t)$ is the subsequent outcome measured after $A(t)$, $t = 1, 2$. One might also observe other changes that have occurred from time $t = 1$ to $t = 2$ which are coded by $E(t)$ at $t = 1, 2$. Thus, the data on one unit i is collected according to the following time-ordering $E(1), W_i(1), A(1), Y_i(1), E(2), W_i(2), A(2), Y_i(2)$. Suppose one wishes to estimate the causal effect of this change in treatment from $A(1) = 0$ to $A(2) = 1$, and let's assume that $A(t)$ can only have two values $\{0, 1\}$.

To formally define a causal effect we define the following NPSEM:

$$\begin{aligned} U &\sim P_U \\ E &= f_E(U_E) \\ W &= f_W(E, U_W) \\ A &= f_A(E, U_A) \\ Y &= f_Y(E, W, A, U_Y). \end{aligned}$$

The observed data corresponds with observing n draws from the counterfactual $(W(e(1)), Y(e(1), a(1)))$ and $(W(e_2), Y(e(2), a(2)))$ corresponding with interventions $E = e(1)$, $A = a(1)$, and $E = e(2)$, $A = a(2)$, respectively. The causal effect of interest is $EY(1) - Y(0)$, where $Y(a)$ is the counterfactual corresponding with intervention $A = a$. Note that $Y(a)$ involves first randomly drawing the environment $E \in \{e(1), e(2)\}$ among the two environments with probability 0.5 on each, drawing covariates from the corresponding environment specific distribution, setting the intervention A at a, and finally drawing the outcome Y. Our identifiability results shows that if (E, W, A) is randomized, and $Y = f_Y(W, A, U_Y)$, then

$$EY(A = 1) - EY(A = 0) = \sum_w \bar{P}_W(w)\{E(Y(2) \mid W(2) = w) - E(Y(1) \mid W(1) = w)\},$$

where $\bar{P}_W(w) = 0.5P_{W(1)}(w) + 0.5P_{W(2)}(w)$. For the sake of estimation of this statistical target parameter, one could still treat this as a two sample problem, as in previous sections. That is, one can represent the data as $2n$ observations on (W, B, Y), $B \in \{(e(1), a(1)), (e(2), a(2))\}$, and apply the targeted MLE for the statistical parameter $E_W E(Y \mid B = (e(2), a(2)), W) - E_W E(Y \mid B = (e(1), a(1)), W)$, treating the sample as $2n$ i.i.d. observations. For statistical inference one now needs to run a bootstrap involving resampling subjects or use influence curve based inference taking into account that the two influence curve values (from i.i.d. influence curve) for the coupled observations on one subject define the single influence curve. That is, our previously presented targeted MLE for the two sample problem is directly applicable, but statistical inference needs to respect the fact that two of the observations are from the same subject.

We now consider a slight variation of the above example in which different subjects are sampled at the second time point. Consider now a study in which we observe a sample of n_1 subjects from a population I in year 1 and observe a sample of n_2 independent subjects from a population II (possibly equal to population I) in year 2. The first sample is exposed to treatment $A(1) = 0$ and the second sample is exposed to treatment $A(2)$. The data on these $n_1 + n_2$ subjects can be coded as $(W_i(t), Y_i(t))$, $i = 1, \ldots, n_t$, $t \in \{1, 2\}$, where $W_i(t)$ denotes the individual history before $A(t)$, and $Y_i(t)$ is the subsequent outcome measured after $A(t)$, $t = 1, 2$. One might also observe other changes in sampling population that have occurred from time $t = 1$ to $t = 2$ which are coded by $E(t)$ at $t = 1, 2$. Thus, the data on one unit i from t-th sample is collected according to the following time-ordering $E(t), W_i(t), A(t), Y_i(t)$, $t = 1, 2$. Suppose one wishes to estimate the causal effect of this change in treatment from $A(1) = 0$ to $A(2) = 1$, and let's assume that $A(t)$ can only have two values $\{0, 1\}$.

To formally define a causal effect we define the following NPSEM:

$$\begin{aligned} U &\sim P_U \\ E &= f_E(U_E) \\ W &= f_W(E, U_W) \\ A &= f_A(E, U_A) \\ Y &= f_Y(E, W, A, U_Y). \end{aligned}$$

We assume that the observed data corresponds with observing n_1 draws from the counterfactual $W(e(1)), Y(e(1), a(1))$ and n_2 draws from $W(e_2), Y(e(2), a(2))$, corresponding with interventions $E = e(1)$, $A = a(1)$, and $E = e(2)$, $A = a(2)$, respectively. The causal effect of interest is $EY(1) - Y(0)$, where $Y(a)$ is the counterfactual corresponding with intervention $A = a$. Note that $Y(a)$ involves first randomly drawing the environment $E \in \{e(1), e(2)\}$ among the two environments with probability $\alpha = n_1/(n_1 + n_2)$ on $e(1)$, drawing covariates from the corresponding environment specific distribution, setting the intervention A at a, and finally

drawing the outcome Y. Our identifiability results shows that if (E, W, A) is randomized, and $Y = f_Y(W, A, U_Y)$, then

$$EY(A=1) - EY(A=0) = \sum_w \bar{P}_W(w)\{E(Y(2) \mid W(2) = w) - E(Y(1) \mid W(1) = w)\},$$

where $\bar{P}_W(w) = \alpha P_{W(1)}(w) + (1-\alpha)P_{W(2)}(w)$. For the sake of estimation of this statistical target parameter, one can treat this as a two sample problem, as in previous sections. That is, one can represent the data as $n_1 + n_2$ i.i.d. observations on (W, B, Y), $B \in \{(e(1), a(1)), (e(2), a(2))\}$, and apply the targeted MLE for the statistical parameter $E_W E(Y \mid B = (e(2), a(2)), W) - E_W E(Y \mid B = (e(1), a(1)), W)$.

5.2 Generalization to multiple time points.

Consider a study in which we observe a sample of subjects from a population over time and expose them to a treatment regimen $A(t)$, $t = 1, \ldots, \tau$. The data on these n subjects can be coded as $(W_i(t), Y_i(t))$, $t \in \{1, \ldots, \tau\}$, $i = 1, \ldots, n$, where $W_i(t)$ denotes the individual history before $A(t)$, and $Y_i(t)$ is the subsequent outcome measured after $A(t)$. Let $E(t)$ denote the environmental factors present at time t and relevant for $Y(t)$, which includes $\bar{A}(t-1) = (A(1), \ldots, A(t-1))$. Thus, the data on one unit i is collected according to the following time-ordering $E(1), W_i(1), A(1), Y_i(1), \ldots, E(\tau), W_i(\tau), A(\tau), Y_i(\tau)$. Suppose one wishes to estimate the causal effect of a change in treatment on the outcome, and let's assume that $A(t)$ can only have two values $\{0, 1\}$.

To formally define a causal effect we define the following NPSEM:

$$\begin{aligned} U &\sim P_U \\ E &= f_E(U_E) \\ W &= f_W(E, U_W) \\ A &= f_A(E, U_A) \\ Y &= f_Y(E, W, A, U_Y). \end{aligned}$$

This NPSEM allows us to define counterfactuals and corresponding causal effects. The causal effect of interest is $EY(1) - Y(0)$, where $Y(a)$ is the counterfactual corresponding with intervention $A = a$, $a \in \{0, 1\}$. Drawing $Y(a)$ involves first randomly drawing the environment $E \in \{e(t) : t = 1, \ldots, \tau\}$ among the t environments with probability $1/\tau$ on each, drawing covariates W from the corresponding environment specific distribution, setting the intervention A at a, and finally drawing the outcome Y. The observed data corresponds with observing n draws from the counterfactual $W(e(t)), Y(e(t), a(t))$ corresponding with interventions $E = e(t)$, $A = a(t)$, across $t = 1, \ldots, \tau$. As before, we reformulate the combined data as an i.i.d. sample on $(B, W(B), Y(B))$ with $B \in \{(e(t), a(t)) : t = 1, \ldots, \tau\}$, and, given $B = (e(t), a(t))$, the distribution of $W(B), Y(B)$ equals the distribution of $W(e(t)), Y(e(t), a(t))$. Our

identifiability results shows that if (E, W, A) is randomized, and $Y = f_Y(W, A, U_Y)$, then, for each t,

$$EY(a) = \sum_w \bar{P}_W(w) E(Y(B) \mid B \in \mathcal{B}_a, W(B) = w),$$

where $\bar{P}_W(w) = \sum_t \frac{1}{\tau} P_{W(t)}(w)$, and $\mathcal{B}_a = \{(e(t), a(t)) : a(t) = a\}$ consists of all time-points for which $a(t) = a$. In particular, this yields the following identifiability result for the additive causal effect of a change in treatment:

$$EY(A = 1) - EY(A = 0) =$$
$$\sum_w \bar{P}_W(w) \left\{ E(Y(B) \mid B \in \mathcal{B}_1, W(B) = w) - E(Y(B) \mid B \in \mathcal{B}_0, W(B) = w) \right\}.$$

For the sake of estimation of this statistical target parameter, one could treat the n observations as a pooled sample of $n * \tau$ observations, ignoring the dependence. That is, one can represent the data as τn observations on (W, B, Y), $B \in \{(e(t), a(t)) : t = 1, \ldots, \tau\}$, and apply the targeted MLE for the statistical parameter $E_W E(Y \mid A = 1, W) - E_W E(Y \mid A = 0, W)$, treating the sample as τn i.i.d. observations, where A denotes the second component of B. For the sake of statistical inference one now needs to run a bootstrap involving resampling from the n subjects, or use influence curve based inference taking into account that the τ influence curve values (from i.i.d. influence curve representation) for the time-series of observations on one subject define the single influence curve. That is, our previously presented targeted MLE for the multi sample problem is directly applicable, but statistical inference needs to respect the fact that the τ observations across time t are from the same subject.

6 Generalization to causal effect of community based intervention on arbitrary parameters of data generating distribution for individuals.

We have been focussing on the causal effect of a community based intervention on the mean outcome. In this section we generalize our approach to causal effect of the community based intervention on arbitrary parameters, thereby including the causal effect of joint interventions at both the community level as well as at the individual level. The next theorem generalizes our identifiability results.

Theorem 5 *Consider a NPSEM,*

$$\begin{aligned} E &= f_E(U_E) \\ W &= f_W(E, U_W) \\ A &= f_A(E, U_A) \\ O &= f_O(E, W, A, U_O). \end{aligned}$$

This allows us to define the counterfactuals $O(e,w,a)$, $O(e,a)$, and $O(a)=O(E,a)$ corresponding with intervention $(E=e,W=w,A=a)$, $(E=e,A=a)$, and $A=a$, respectively. Let $\alpha()$ be the marginal distribution of E on $\{e_1,\ldots,e_J\}$.

We observe n_j i.i.d. observations on $(W(e_j),O(e_j,a_j))$ for a collection of $(e_j,a_j)\in\mathcal{B}=\{(e_1,a_1),\ldots,(e_J,a_J)\}$, $j=1,\ldots,J$. Let P_b be the distribution of $(W(b),O(b)$ for $b\in\mathcal{B}$.

We reformulate this observed data set as $n=\sum_j n_j$ i.i.d. on $(B,W(B),O(B))$, where $P(B=(e_j,a_j))=n_j/n$ and $P(W(B),O(B)\mid B=b)\sim P_b$.

We assume the strong randomization assumption stating that (E,W,A) is independent of $O(e,a,w)$ for all (e,a,w), and the exclusion restriction assumption $O=f_O(W,A,U_O)$ stating that O is not a function of E.

We have the following identifiability result:

$$P(W(E)=w,O(E,a)=o)=\bar{P}(w)P(O(B)=o\mid B\in\mathcal{B}_a,W=w),$$

where $\bar{P}(w)=\sum_{b=(e,a)}P(W(e)=w)\alpha(e)$. In particular,

$$P(O(E,a)=o)=\sum_w\bar{P}(w)P(O(B)=o\mid B\in\mathcal{B}_a,W=w).$$

Proof. For notational convenience, we denote $f_O(W,A,U_O)$ with $f(W,A,U)$. We have

$$\begin{aligned}
&P(W(E)=w,O(E,a)=o)=P(W(E)=w,f(W,a,U)=o)\\
&=\textstyle\sum_e P(E=e,W(e)=w,f(w,a,U)=o)\\
&=\textstyle\sum_e P(O(a,w)=o\mid W(e)=w,E=e)P(W(e)=w,E=e)\\
&=\textstyle\sum_e P(O(a,w)=o)P(W(e)=w,E=e)\\
&\text{ by strong RA}\\
&=P(O(a,w)=o)\textstyle\sum_e P(W(e)=w,E=e)\\
&=P(O(a,w)=o)\bar{P}(w)\\
&\text{ by definition of }\bar{P}(w)\\
&=P(O(a,w)=o\mid W(e)=w)\bar{P}(w)\\
&\text{ by strong RA.}\\
&=P(O(B)=o\mid B=(e,a),W(e)=w)\bar{P}(w),
\end{aligned}$$

where we used at last step that

$$P(O(a,w)=o,W(E)=w)=P(O(B)=o,W(E)=w\mid B=b),$$

and thereby

$$P(O(a,w)=0\mid W(E)=w)=P(O(B)=o\mid B=b,W(E)=w).$$

Let $A(B)$ denote the A-component of $B=(E,A)$. We have for any function h,

$$\begin{aligned}
&E(h(O(B))\mid A(B)=a,W(B)=w)=E\{E(h(O(B))\mid B,A(B)=a,W(B)=w)\mid A(B)=a,W(B)=w\}\\
&=\{E(h(O(B))\mid B,A(B)=a,W(B)=w)\},
\end{aligned}$$

where we used that $E(h(O(B)) \mid B, W(B) = w)$ only depends on B through A. Thus, by letting $h(o) = I(O = o)$, it follows

$$P(O(B) = o \mid B = (e, a), W(B) = w) = P(O(B) = o \mid B \in \mathcal{B}_a, W(B) = w).$$

This proves now

$$P(W(E) = w, O(E, a) = o) = P(O(B) = o \mid B \in \mathcal{B}_a, W(B) = w)\bar{P}(w).$$

This completes the proof. □

Identifiability of causal parameters. Suppose one wishes to estimate a particular parameter of the distributions of $O(E, a) \sim P_a$ for different a, such as

$$\Psi(P_1) - \Psi(P_0).$$

For example, $\Psi(P_a)$ might represent a parameter of a G-computation formula for the counterfactual distribution of $O(E, a)$ under an intervention on an individually measured treatment A^I that is included in the observation $O(E, a)$: i.e., this would be a counterfactual $O(E, a, a^I)$, defined in an augmented NPSEM, corresponding with a joint intervention on A and A^I. In this manner, $\Psi(P_a)$ corresponds with a distribution corresponding with setting a community based intervention $A = a$ and an individual treatment regimen $A^I = a^I$.

Under the stated no residual environmental confounding assumption (exclusion restriction assumption on NPSEM), the theorem teaches us that we can identify P_a as the distribution of O under the density $P_{O(B)|B\in\mathcal{B}_a,W(B)}P_{W(B)}$ for a joint (O, W). For many parameters Ψ

$$\Psi(P_a) = E_{W(B)}\Psi(P_{O(B)|B\in\mathcal{B}_a,W(B)}) \equiv \Psi_a(P),$$

i.e., one can evaluate it as an average over w w.r.t distribution of $W(B)$ of the same parameter of the conditional distribution of $O(B)$, given $B \in \mathcal{B}_a, W(B) = w$. In general, $P_a = P_a(P)$ is now identified by the distribution P of observed data structure $(B, W(B), O(B))$.

G-computation formulas for joint community and individual interventions. The above theorem states that $P(W(E) = w, O(a) = o) = P(W(E) = w)P(O(B) \mid A = a, W(E) = w)$. This teaches us also that the distribution of $O(a, a^I)$ under a joint intervention including an intervention $A^I = a^I$ in an augmented NPSEM (incorporating equations for O), is the same as the G-computation formula for the joint intervention treating the data as $(W, A, (A^I, L(A, A^I))$, i.e., respecting the time-ordering, ignoring E, and treating the data as observed directly from the NPSEM. This means that we can now obtain, under the stated conditions, identifiability results for all wished causal effects of joint community and individual treatment assignments.

The statistical estimation problem. Thus, the above theorem establishes the wished identifiability result representing the wished causal parameter as a parameter $\Psi(P)$ of the probability distribution of the observed data structure $(B, W(B), O(B))$. Suppose that the model $\mathcal{M}$ for P is the nonparametric model: we prefer not to include the no-residual confounding assumption in the observed data model, since we are interested in statistical inference for this same parameter $\Psi(P)$ without this assumption as well, in particular, incorporating the residual environmental confounding in an variance estimate. The statistical estimation problem is now defined: we observe i.i.d. $(B_i, W_i(B_i), O_i(B_i)) \sim P$, $i = 1, \ldots, n$, $P \in \mathcal{M}$, and we wish to estimate the statistical parameter $\Psi(P)$ such as $\Psi_a(P)$ or a contrast $\Psi(P)$ representing $\Psi_1(P) - \Psi_0(P)$.

Double robust mapping for identifying mean of functions of counterfactuals. To construct a semi-parametric efficient estimator of $\Psi(P)$, such as the targeted MLE, one will need to obtain the efficient influence curve $D^*(P)$ of Ψ. We are now concerned with presenting a general approach for obtaining this efficient influence curve, assuming that it is well understood how to obtain the efficient influence curve of $\Psi(P)$ if one would have observed directly from the NPSEM (W, A, O).

The identifiability theorem for the distribution of $O(a)$ has the following implications for identifying a mean of a function $D(O(a))$, which could represent an estimating function or loss function.

Lemma 1 *Consider a function D of $O(a)$. Under the assumptions stated in previous theorem, including the exclusion restriction assumption, the following holds.*

We have the following Inverse probability of community intervention mapping:

$$ED(O(a)) = E\left\{D(O(B))\frac{I(B \in \mathcal{B}_a)}{P(B \in \mathcal{B}_a \mid W(B))}\right\}.$$

We have the following double robust inverse probability of community of intervention mapping:

$$\begin{aligned} S(D) &= \frac{I(B \in \mathcal{B}_a)}{P(B \in \mathcal{B}_a \mid W(B))}\{D(O) - E(D(O) \mid B \in \mathcal{B}_a, W(B))\} \\ &\quad + E(D(O) \mid B \in \mathcal{B}_a, W(B)). \end{aligned}$$

View $S = S(Q, g, D)$ as indexed by nuisance parameters $g_0(a \mid W(B)) = P(B \in \mathcal{B}_a \mid W(B))$, and $Q_0(a, W(B)) = E(D \mid B \in \mathcal{B}_a, W(B))$. We have

$$E(S_{Q,g}(D)) = ED(O(a)),$$

if $g(a \mid W(B)) > 0$ a.e., and either $Q = Q_0$ or $g = g_0$.

This mapping can be used to map an (optimal and double robust) estimating function or loss function (e.g., loglikelihood) based on sampling i.i.d. $O(a)$ into an (optimal and double robust) estimating function based on the observed data structure $(B, W(B), O(B))$, where B itself is reduced to just the A component.

General algorithm for computing the efficient influence curve. Given our assumptions, we have the following general strategy for computing the efficient influence curve of a causal parameter $\Psi^F((P_a : a)) = \Psi(P)$, which can be viewed as a function of the counterfactual distributions P_a of $O(a) = O(E, a)$ for one or more a-values.

- View the observed data structure $(B, W(B), O(B))$ as $(W, A, O(A))$, ignoring E, where $O(a) \sim P_a$. Thus our observed data is viewed as a standard point treatment missing data structure on counterfactuals $O(a)$ and the covariates W play the role of baseline/pre-treatment covariates. In addition, treat A as being randomized, conditional on W: $P(A = a \mid W, O(a)) = P(A = a \mid W)$. In this world we can identify $P(W = w, O(a) = o) = P(O(A) \mid A = a, W = w)P(W = w)$ with the standard point-treatment G-computation formula, which corresponds exactly with our identifiability result. Use this G-computation formula to represent $\Psi^F(P_a : a) = \Psi(P)$ as parameter of the observed data distribution of $(W, A, O(A))$, which again, corresponds exactly with our identifiability result.

- In this standard missing data problem, compute the efficient influence curve $D^*(P)$ for the parameter $\Psi^F(P_a : a) = \Psi(P)$ at a P of $(W, A, O(A))$. This will also be the efficient influence curve based on the actual data structure $(B, W(B), O(B))$, suppressing information of E.

Remark. We suggest that E could be included in $W(B)$, and the algorithm used to construct the estimator of $\Psi(P)$ needs to decide if the particular E factors are resulting in too much violation of $P(A = a \mid W(B)) > 0$ a.e. In this way, if the number of communities grows, certain factors of E can start to be included and adjusted for. In particular, the collaborative targeted maximum likelihood estimator could be used to data adaptively decide which E factors to still include.

6.1 Example: Causal effect of combined community based intervention and individual treatment.

Suppose that we have two communities, one gets assigned a treatment and another a control. We sample individuals from the two communities, and on each individual we observe the data structure (W, W^I, A^I, Y), where W are pre-community intervention baseline covariates, W^I are pre-treatment covariates, A^I is an individually assigned treatment, and Y is an outcome of interest. We assume an NPSEM for the nodes

E, A, W, W^I, A^I, Y, allowing us to define counterfactuals of Y under set values of E, A, W. It is assumed that the functions for W^I, A^I, Y exclude E: i.e., we assume that the effect of E on the individual data structure is blocked by the covariate W. We also assume the strong randomization assumption stating that (E, A, W) is independent of $(W^I, A^I, Y)(e, a, w)$ for all set values of (e, a, w). Suppose we wish to estimate the mean counterfactual outcome of $EY(a, a^I)$ under a set community based intervention $A = a$ and a treatment $A^I = a^I$. We can represent the two samples as a combined i.id. sample on $(B, W(B), O(B))$, where $B = (A, E) \in \{(1, e_0), (0, e_1)\}$, and $O(B) = (W^I(B), A^I(B), Y(B))$.

Following the general recipe presented above, to estimate the counterfactual mean we can treat the observed data structure as i.i.d. observations on (W, A, W^I, A^I, Y) and identify the distribution of $(W, W^I, Y)(a, a^I)$ by the G-computation formula:

$$P(W)P(W^I \mid A = a, W)P(Y \mid A^I = a^I, W^I, A = a, W),$$

which also identifies the marginal distribution of $Y(a, a^I)$. In particular, we can estimate $EY(a, a^I)$ with the targeted MLE of $EY(a, a^I)$ defined as this parameter of the G-computation formula.

7 Practical conclusion.

What have we learned after this journey?

Identifying and estimating a causal effect of treatment on an outcome distribution based on data generated by assigning a treatment at the unit-level, possibly in response to baseline and intermediate covariates, across many units, is by now an intensively studied and reasonably well understood problem: that is, one observes n i.i.d. copies $(A_i, X_i(A_i))$, and one wishes to estimate a parameter of the distribution of $X(a)$ for some specified a-values under the assumption that the treatment assignment is a deterministic function of observables (only), and a non-zero exogenous error (so that there is treatment experimentation given all these observed confounder values). In this case the fundamental problem to address is to utilize covariates measured at individual level to control for the fact that the treatment empirically or theoretically is a function of such covariates. The identification problem is addressed by the G-computation formula under the sequential randomization assumption, and, for example, semiparametric model-based efficient targeted maximum likelihood estimators of the resulting statistical target parameter of the distribution of the data have been developed.

A different problem is to identify and estimate a causal effect of treatment on an outcome distribution based on data generated by assigning a single treatment to a community of units, across few communities that will differ by environmental factors. That is, one observes many observations on a counterfactual $X(e_j, a_j)$, across a few values (e_j, a_j) of environmental settings e_j and treatment value a_j, $j = 1, \ldots, J$. The fundamental problem to address is to utilize covariates measured

at the individual level to control for the fact that the community that was exposed to treatment had a different environment than the community that was exposed to control. Stating that treatment level was assigned randomly (or deterministically in balanced way) to the few communities is good, but far from sufficient. Stating that one measures the environmental factors that make the communities different is good, since it allows some adjustment by them, but is far from sufficient, and useless if there are say only two communities.

The identification result for a causal effect of treatment is now obtained by assuming that the outcome does not depend on the environmental factors beyond the measured pre-treatment covariates, and one also needs to assume that the these pre-treatment covariates are randomized itself, given the environmental factors. These assumptions allow us to identify the counterfactual outcome distribution corresponding with setting the treatment in the NPSEM.

Given the identification result, the statistical parameter that identifies the wished causal effect of treatment is defined, so that the statistical estimation problem is well defined. We presented a targeted MLE of the counterfactual mean $EY(a)$, which is a substitution estimator obtained by plugging in an estimator of $\bar{Q}_0(w,a) = E(Y(B) \mid W(B) = w, B \in \mathcal{B}_a)$, and then averaging w.r.t. the empirical distribution of the pooled sample of the covariates. That is, we can treat the data structure as (W, A, Y) and apply the targeted MLE for estimation of the target parameter $E_W E(Y \mid W, A = a)$. This is what a naive person would have done who is ignorant of the underlying data generating experiment, but treats it as the regular causal inference problem in which each individual gets assigned a treatment A in response to the baseline covariates W.

We showed that without the exclusion restriction assumption, the targeted statistical parameter will still be a (non-wished) causal effect, but the latter is subject to bias w.r.t. a wished causal effect, due to a difference between the empirical distribution of the environmental factors in the treated and untreated communities. We show that this bias will disappear when the number of communities increases and the treatment is randomized across communities. We present a variance estimate that takes into account the residual bias, thereby allowing us to do honest statistical inference for the wished causal effect in the model that does not assume the exclusion restriction assumption, while our variance estimate still approximates the uncorrected variance if the exclusion restriction assumption happens to hold, and, always adapts to the degree at which the exclusion restriction assumptions fails to hold.

In addition, we found that this approach to causal effect estimation of a community based intervention, possibly combined with an intervention on individually assigned treatment nodes, can be completely generalized: for each community, for each individual in that community, determine the pre-intervention covariates W that might block the effect of the environment on the outcome, add the community based intervention A as a treatment that comes after W, augment this (W, A) with

the subsequent longitudinal data structure measured on the individual that might also include individually assigned treatment nodes, and proceed as if the goal is to estimate a causal effect of an intervention on A and possibly other nodes in the longitudinal data structure, using the standard G-computation formula, and corresponding targeted MLE. That is, the estimators developed for estimation of causal effects of individually assigned interventions based on an NPSEM, the consistency assumption, and sequential randomization assumption, can be applied to estimate the effects of community based interventions combined with interventions on individually assigned treatments based on community based sampling, by formulating the data on each unit as (W, A, O) and proceed as usual viewing A as an initial treatment node. The assumptions under which these statistical target parameters represent the wished causal effect now include, beyond the sequential randomization assumption needed for the individually assigned treatment, the assumption that W blocks the effect of the environment on O, beyond randomization (E, W, A) in the underlying NPSEM.

8 Handling dependence among sampled individuals within a community.

Suppose that the NPSEM proposed in previous sections applies for each marginal draw with a common marginal distribution for the exogenous input U, but inputs U of the NPSEM are correlated among individuals in the community: i.e., given a draw of E, A, the repeated draws of W, Y are correlated, but, marginally, the NPSEM applies. We define a causal effect as a parameter of this common marginal distribution of (U, X). Our observed data sample is still $n = n_0 + n_1$ draws of $(B, W(B), Y(B))$, representing the two-sample study corresponding with a treatment and control region, but, given $B = (1, e_1)$, the repeated draws are correlated, and similarly, given $B = (0, e_0)$.

For example, suppose our target parameter is still the additive causal effect $EY(A = 1) - EY(A = 0)$, or the causal effect among the treated. This target parameter is only a function of the common marginal distribution of the U's: it is not affected by the joint distribution of the U's across the individuals in the community. The identifiability result also still applies, since it is a statement about writing a parameter of the common distribution of (U, X) as a function of the corresponding distribution of O, so that such an identifiability result is not affected by the joint distribution of the U's. Thus, this result show, under the exclusion and randomization assumption, $EY(1) - Y(0) = E_{W(B)}\{E(Y(B) \mid B = (1, e_1), W(B)) - E(Y(B) \mid B = (0, e_0), W(B))\}$, where the parameter of the observed data distribution of $(B, W(B), Y(B))$ only concerns the common marginal distribution of $(W_i(1, e_1), Y_i(1, e_1))$, $i = 1, \ldots, n_1$, and the common marginal distribution of $(W_i(0, e_0), Y_i(0, e_0))$, $i = 1, \ldots, n_0$. That is, we are not concerned with a target parameter that depends on this joint distribution of the U's across the units.

This now suggests to estimate this target parameter with the same targeted maximum likelihood estimator treating the sample as n i.i.d. observations $(B_i, W_i(B_i), Y_i(B_i))$, $i = 1, \ldots, n = n_0 + n_1$. This is not different from using a pooled estimator treating the observations as i.i.d to a repeated measures regression.

Incorporating dependence when estimating the standard error. It is at the point of assessment of the uncertainty/standard error of this targeted MLE of ψ_0 that we need to take into account the dependence between the units. Can we provide an inferential method that provides a reasonable adjustment while it still reduces to the i.i.d statistical inference when the dependence is negligible? Since the estimator is algorithmically identical as applied to the i.i.d. case we will assume that the same first order Tailor expansion is appropriate to base inference on. Let $IC(O_i)$ be the influence curve of the targeted MLE ψ_n. So we will work with the approximation that

$$\psi_n - \psi_0 \approx (P_n - P_0)IC = \frac{1}{n}\sum_{i=1}^{n} IC(O_i). \tag{6}$$

We will assume that the dependence between individuals is weak enough so that ψ_n is still asymptotically normally distributed. In particular, it is assumed that $(\psi_n - \psi_0)/SE(\psi_n)$ converges to a normal distribution with mean zero and variance 1, where $SE(\psi_n)$ is the standard error of ψ_n. We will estimate the variance of ψ_n with the variance of $1/n\sum_i IC(O_i)$, where this variance can be decomposed as

$$\sigma_n^2 = \frac{1}{n}\sum_{i=1}^{n} EIC(O_i)^2 + \frac{2}{n^2}\sum_{i<j} E\{IC(O_i)IC(O_j)\}.$$

Under independence the second term equals zero, but in this case we will estimate this contribution, which results in the estimate

$$\hat{\sigma}_n^2 = \frac{1}{n}\sum_{i=1}^{n} \hat{IC}(O_i)^2 + \frac{2}{n^2}\sum_{i<j} \hat{IC}(O_i)\hat{IC}(O_j).$$

We note that this estimate will be asymptotically equivalent with the estimate $1/n\sum_i \hat{IC}^2(O_i)$ one would use in the i.i.d. case, if there happens to be no dependence. In addition, if in truth the n observations consist of m i.i.d. clusters of J observations, but this is unknown to us, then

$$\hat{\sigma}_n^2 \approx \frac{1}{m}\sum_{i=1}^{m}\left\{\frac{1}{J}\sum_{j=1}^{J} IC(O_i(j))\right\}^2.$$

That is, in this case the estimate of the variance corresponds with respecting the fact that the true influence curve of the estimator is $\bar{IC}(O_i) = 1/J\sum_j IC(O_i(j))$ for the

cluster observation $O_i = (O_i(j) : j)$, and thereby obtains the right variance of the estimator. In general, this estimator does not require knowledge of the dependence structure and thus also handles the case that there is no replication of independent units, as long as the dependence is weak enough so that a CLT applies. Statistical inference is now based on the approximation that $\psi_n \sim N(\psi_0, \hat{\sigma}_n^2)$. For example, a 0.95-confidence interval would be $\psi_n \pm 1.96\hat{\sigma}_n$.

Diagnosing too much dependence for a CLT-based confidence interval. Even though the above method will be able to adapt to underlying (but unknown) cluster dependence it does rely on the assumption that $\psi_n - \psi_0$, and, in particular, its first order expansion $1/n \sum_i IC(O_i)$, is asymptotically normally distributed. Thus, the amount of dependence has to be limited enough so that a CLT-approximation is still valid. Having presented the above method, we would now like to provide a tool to diagnose that the data does not allow a CLT-based approximation of $\psi_n - \psi_0$.

We wish to investigate if $1/n \sum_i IC(O_i)$ follows approximately a mean zero normal distribution. For that purpose we propose a resampling method for resampling $O_1^\#, \ldots, O_n^\#$ for which the marginal distribution of $O_j^\#$ is the empirical P_n, as in the regular non-parametric bootstrap, but for which the joint distribution of $O_1^\#, \ldots, O_n^\#$ is such that

$$\text{VAR}\left(\frac{1}{n}\sum_i IC(O_i^\#\right) = \hat{\sigma}_n^2.$$

In other words, we set the dependence level so that it corresponds with our estimate of the variance of the linear approximation of $\psi_n - \psi_0$. Under such a sampling distribution we can now evaluate the distribution of $\bar{IC}_n/\hat{\sigma}_n$, where $\bar{IC}_n = 1/n \sum_i IC(O_i)$, and, the distribution of $(\psi_n^\# - \psi_n)/\hat{\sigma}_n$, and determine if these two distributions are indeed approximately $N(0, 1)$.

We have two particular proposals. Firstly, one could come up with some model for a joint distribution of $O_1, \ldots, O_n$ incorporating dependence with some tuning parameter α, simulate a large number B (e.g., 10,000) times a correlated set of n observations, $O_b^\# = (O_{1b}^\#, \ldots, O_{nb}^\#)$, $b = 1, \ldots, B$, and corresponding influence curves $IC(O_{1b}^\#), \ldots, IC(O_{nb}^\#)$. In this model one does not necessarily worry about the marginals of $O^\#$ being equal to the empirical distribution, since we correct for this in the next step.

Let $F_1^\#, \ldots, F_n^\#$ be the marginal cumulative distribution functions of $IC(O_{1b}^\#), \ldots, IC(O_{nb}^\#)$, respectively. Let F_n be the empirical cumulative distribution function of $IC(O_1), \ldots, IC(O_n)$. Let $Q_j = F_n^{-1} F_j^\#$ be the quantile-quantile function that maps a random variable with distribution $F_j^\#$ into a random variable with distribution F_n. This quantile-quantile function can be generalized to discrete distributions as in Yu and van der Laan (2002): $x \to F_n^{-1}(UF(x) + (1-U)F(x-))$, where $U \sim U(0, 1)$, is the generalized quantile-quantile function mapping a random variable with distribution F (e.g, discrete) into random variable with distribution F_n. We now consider the trans-

formed $IC_{1b}^{\#} = Q_1(IC(O_{1b}^{\#})), \ldots, IC_{nb}^{\#} = Q_n(IC(O_{nb}^{\#}))$ resampled influence curves, $b = 1, \ldots, B$. These still reflect the dependence structure of the original resampled influence curves, but the marginal distributions are now equal to the empirical F_n as required. The tuning parameter α is now fine-tuned to obtain the level $\hat{\sigma}_n^2$ for the variance of $1/n \sum_i IC_i^{\#}$ across its B-replicates.

Secondly, concretely, we propose the following nonparametric bootstrap method for creating the wished dependence while maintaining the marginal distributions equal to the empirical. From $k = 1, \ldots, K$ sample K i.i.d. observations $O_k^{\#}$ from the empirical distribution P_n. From $k = K+1, \ldots, n$, sample $n - K$ i.i.d. observations $O_k^{\#}$ by drawing from the uniform distribution on $O_1^{\#}, \ldots, O_K^{\#}$. This results in a sample $O_1^{\#}, \ldots, O_n^{\#}$ whose marginal distributions are P_n, but the effective sample size is $K < n$, thereby creating dependence. The smaller one chooses K the more dependence one incorporates and $K = n$ corresponds with full independence. One selects K so that the variance of $1/n \sum_i IC(O_i^{\#})$ across the B replicates equals $\hat{\sigma}_n^2$. For this choice of K we evaluate the distribution of $\bar{IC}_n/\hat{\sigma}_n$, where $\bar{IC}_n = 1/n \sum_i IC(O_i)$, and, the distribution of $(\psi_n^{\#} - \psi_n)/\hat{\sigma}_n$, and determine if these two distributions are indeed approximately $N(0, 1)$.

Remark about interactivity of individuals modifying the treatment effect. In many cases, the intervention assigned to a community affects the individual outcomes not only directly but also indirectly through other individuals that interact with the individual. The NPSEM for the marginal distribution for a randomly drawn individual from such a community could involve covariate measurements in W that measure the interactivity of the individual with others in the community. In this way, the marginal distribution modeled by the NPSEM would thus still model enhanced or suppressed effects of treatment for heavily connected individuals relative to less connected individuals. In addition, our proposed standard error estimate will take into account the effect of dependence of the exogenous errors/inputs U of the NPSEM on the standard error of the targeted MLE estimator of the target causal effect defined as a parameter of the marginal distribution of the NPSEM, without a need to specify elaborate random effect or other dependence models for which we lack knowledge.

9 Discussion.

This article provided a number of non-obvious contributions to the literature on causal inference for community based interventions.

In the setting that we observe two populations under two different treatment regimens, while collecting data at the individual level, we made the following contributions. We define a causal effect of treatment intervention a_1 versus a_0 as the difference in means between treatment group and control group in the ideal experiment in which one is able to randomize each unit of the combined population to

treatment or control. We show that this causal effect can be identified under a randomization assumption and under the assumption that we can collect covariates at the individual level that are not affected by the treatment and that block the effect of the differences in the environment between the two populations. As a consequence, for this target parameter there is a clear role for covariates, but different in flavor from the classical causal inference case in which treatment is assigned at individual level: the main purpose of the covariates is to remove bias by blocking the effect of different environments between the populations on the outcome distributions.

Moreover, the statistical parameter representing this additive causal effect involves computing the mean outcome as a function of the covariates for the treatment population, and similarly for the control population, taking the difference between these covariate-value specific mean outcomes, and averaging the difference over all units in the combined sample.

Secondly, we present the (collaborative) targeted maximum likelihood estimator, based on first using super learning to estimate these regressions, and a subsequent targeted maximum likelihood step relying on an estimate of the probability of being selected in population 1 as a function of the covariates. The targeted maximum likelihood estimator is double robust and efficient. That is, with our formulation we make it possible to use the state of the art statistical methodology in causal inference, and obtain fully efficient and double robust estimators of the causal effect of an intervention assigned at the population.

Thirdly, we show that this approach, somewhat surprisingly, can also be used to estimate the additive causal effect of setting treatment at time t (choosing between the observed treatment level at time t for population 1 versus the observed treatment level at time t for population 2). The past treatment and past environment is viewed as the environment variable assigned at the population level, current treatment at time t is viewed as the treatment assigned at the population level, and the individual past before treatment at time t represents the covariates that block the effect of differential environment. In this manner, the same methodology can be applied, providing us with double robust and efficient targeted maximum likelihood estimators of the t-specific causal effects of community based interventions, and user-supplied summary measures of these t-specific causal effects.

Another contribution concerns the extension to matched cohort designs for these studies. We extend our estimators to matched cohort sampling in which individuals from the treated population are matched to one or more individuals from the control population. We use general results on efficient influence curves and targeted maximum likelihood estimation for case-control sampling as established in van der Laan (2008) to compute the semiparametric information bounds and present the targeted maximum likelihood estimator for these matched cohort designs. We show that the matched cohort design is very much targeted towards the causal effect among the treatment population, showing the strong potential benefits of matching for the purpose of this causal effect among the treated.

Up till this point we focussed on the case that the number of communities is small (two), but, the estimators are perfectly applicable to the case that the number of communities is large, which is in a sense a less challenging case. In particular, we generalized our identifiability theorems to the case that one assigns two possible interventions to J communities. We presented the efficient influence curve for the target parameter and the targeted MLE. In particular, our i.i.d. representation of this multi-sample data structure in terms of $(B, W(B), Y(B))$ shows that the effective sample size will be $\sum_j n_j$, the sample size across all communities.

Finally, we considered the case that the exclusion restriction assumption does not hold. We use the same targeted MLE if the exclusion assumption is assumed to hold or not. We show that our statistical target parameter (as estimated by the targeted MLE) remains a well understood causal effect (involving adjustment by the environments as well), but one that is subject to residual bias due to a difference in the empirical distribution of the environmental factors between the treated and untreated communities. We redefine a causal target that is unconfounded by the environmental factors by averaging across an infinite sample of environments/communities, while it still equals the additive causal effect of treatment, $EY(1) - Y(0)$, under the exclusion restriction assumption, for each fixed number of communities. Asymptotics of the targeted MLE relative to this new generalized causal target in both the number of communities as well as sample sizes within communities is used to establish a central limit theorem for the targeted MLE minus this generalized causal effect. The bias due to the residual environmental confounding is viewed as a mean zero random variable, so that it naturally translates into a augmented variance estimate. This results in an adaptive variance estimator that naturally adapts to the degree of violation of the exclusion restriction assumption, and the number of communities J: so even for J finite, it can result in a variance estimate comparable with the variance of the targeted MLE if the exclusion restriction assumption happens to hold, but a variance that is $O(1/J)$ represents the other extreme possibility.

Subsequently, we noted that the identification of the causal effect for sampling J communities that are different due to environmental factors E can be mapped into an identification of the causal effect of sampling individuals at different time-points: i.e., let time t play the role of E. In this manner, we extended our results to identification and estimation of a causal effect of treatment at time t based on following up a single cohort of individuals exposed to a single time-dependent treatment, under a same exclusion restriction assumption on an appropriate NPSEM that allows one to define the causal effect.

In many community based intervention studies there is dependence between the sampled units within the communities due to interaction of the units. A measure of interconnectivity of an individual should be an important covariate to measure and to utilize, allowing the NPSEM to define causal effects that are modified by the interconnectivity of the unit. If the community is the experimental unit and one samples many communities, then the i.i.d. causal inference estimation methodology

and theory is applicable. If the number of communities is small and there is dependence among individuals within the community, then the only hope for statistical inference is that the dependence is weak enough to allow for a central limit theorem. For that purpose we provided an adjustment in statistical inference that is able to estimate the standard error of our estimate of the causal effect that takes into account hidden dependencies, such as underlying cluster structure. The observed dependence, such as the ratio between an i.i.d. based variance estimate and our proposed variance estimate that incorporates correlations, is an interesting parameter itself, providing a measure of the effective sample size each community-specific sample provides. This might be important designing studies w.r.t. power.

Our proposed approach to handle dependence needs to be evaluated in simulation studies and studied in more detail. Finally, we generalized our findings to arbitrary data structures measured at the individual level, and estimation of causal effect of the community based intervention combined with interventions on treatment or censoring nodes for the individually collected data structure. With this generalization, our work provides a general recipe for analyzing complex observational studies that also involve community based interventions.

References

M. Aitkin and N. Longford. Statistical modeling issues in school effectiveness studies. *Journal of the Royal Statistical Society, Series A*, 149:143, 1986.

D.A. Barr. The effects of organizational structure on primary care outcomes under managed care. *Annals of Internal Medicine*, 122:353359, 1995.

L.F. Berkman, T. Glass, I. Brissette, and T.E. Seeman. From social integration to health: Durkheim in the new millennium. *Social Science & Medicine*, 51:843857, 2000.

A. Biglan, K. Smolkowski, T. Duncan, and C. Black. A randomised controlled trial of a community intervention to prevent adolescent tobacco use. *Tobacco Control*, 9:2432, 2000.

A.S. Bryk and S.W. Raudenbush. *Hierarchical linear models*. Newbury Park: Sage, 1992.

D.T. Campbell and J.C. Stanley. *Experimental and quasi-experimental designs for research*. Hopewell, NJ: Houghton Mifin, 1963.

J. Cassel. The contribution of the social environment to host resistance. *American Journal of Epidemiology*, 104:107–123, 1976.

J.R. Charlton, M.F. DSouza, M. Tooley, and R. Silver. A community trial strategy for evaluating treatment for symptomatic conditions. *Statistics in Medicine*, 4: 1121, 1985.

J. Clark, D.A. Potter, and J.B. McKinlay. Bringing social structure back into clinical decision making. *Social Science & Medicine*, 32:853866, 1991.

C.C. Clogg and A. Haritou. The regression method for causal inference and a dilemma confronting this method. In *In S. P. Turner (Ed.), Causality in crisis? Statistical methods and the search for causal knowledge in the social sciences*, page 83112. Notre Dame, IN: Notre Dame University Press, 1997.

W.G. Cochran. The planning of observational studies of human populations. *Journal of the Royal Statistical Society, Series A*, 128:243265, 1965.

J.S. Coleman, E.Q. Campbell, C.J. Hobson, J. McPartland, A.M. Mood, F.D. Weinfeld, and R.L. York. Equality of educational opportunity. washington, dc: Government printing ofce. *Journal of the Royal Statistical Society, Series A*, 156:938, 1966.

J.B. Copas and H.G. Li. Inference for non-random samples (with discussion). *Journal of the Royal Statistical Society, Series B*, 59:5595, 1997.

A.P. Dempster, N.M. Laird, and D.B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society, Series B*, 39: 18, 1977.

D. Draper. Inference and hierarchical modeling in the social sciences. *Journal of Educational and Behavioral Statistics*, 20:115147, 1995.

H.A. Feldman, J. B. McKinlay, D.A. Potter, K.M. Freund, R.B. Burns, M.A. Moskowitz, and L.E. Kasten. Nonmedical inuences on medical decision making: An experimental technique using videotapes, factorial design, and survey sampling. *Health Services Research*, 32:343366, 1997.

H.A. Feldman, M.A. Proschan, D.M. Murray, D.C. Goff, M. Stylianou, E. Dulberg, P.G. McGovern, W. Chan, N.C. Mann, and V. Bittner. Statistical design of react (rapid early action for coronary treatment), a multisite community trial with continual data collection. *Controlled Clinicial Trials*, 19:391403, 1998.

D. Freedman. From association to causation via regression (with comments). in s. p. turner (ed.). 1997.

H. Goldstein. *Multilevel statistical models in educational and social research.* London: Edward Arnold, 1987.

H. Goldstein. *Multilevel statistical models.* London: Edward Arnold, 1995.

S. Greenland. Ecologic versus individual-level sources of bias in ecologic estimates of contextual health effects. *International Journal of Epidemiology*, 30:13431350, 2001.

S. Greenland. A review of multilevel theory for ecological analyses. *Statistics in Medicine*, 21:389395, 2002.

S. Greenland. Randomization, statistics, and causal inference. *Epidemiology*, 1: 421429, 1990.

S. Gruber and M.J. van der Laan. An application of collaborative targeted maximum likelihood estimation in causal inference and genomics. *The International Journal of Biostatistics*, 2010.

M.E. Halloran and C.J. Struchiner. Causal inference in infectious diseases. *Epidemiology*, 6:142151, 1995.

J.J. Heckman. Sample selection bias as a specication error. *Econometrica*, 47: 153162, 1979.

H.D. Holder, R.F. Saltz, A.J. Treno, J.W. Grube, and R.B. Voas. Evaluation design for a community prevention trial. an environmental approach to reduce alcohol-involved trauma. *Evaluation Review*, 21:140165, 1997.

E.B. Hook. (letter to editor) re: Neighborhood social environment and risk of death: Multilevel evidence from the alameda county study. *American Journal of Epidemiology*, 151:11321133, 2001.

J. S. Kaufman and S. Kaufman. Assessment of structured socioeconomic effects on health. *Epidemiology*, 12:157167, 2001.

J.S. Kaufman and R.S. Cooper. Seeking causal explanations in social epidemiology. *American Journal of Epidemiology*, 150:113120, 1995.

J.S. Kaufman and C. Poole. Looking back on causal thinking in the health sciences. *Annual Review of Public Health*, 21:101119, 2000.

I.G.G. Kreft, J. de Leeuw, and R. van der Leeden. Review of ve multilevel analysis programs: Bmdp-5, genmod, hlm, ml3, varcl. *American Statistician*, 48:324335, 1994.

N. Krieger. Theories for social epidemiology in the 21st century: An ecosocial perspective. *International Journal of Epidemiology*, 30:668677, 2001.

N.M. Laird and J.H. Ware. Random-effects models for longitudinal data. *Biometrics*, 38:963974, 1982.

E. Leamer. Lets take the con out of econometrics. *American Economic Review*, 73: 3243, 1983.

J. De Leeuw and I.G.G. Kreft. Software for multilevel analysis. In *In H. Goldstein (Ed.), Multilevel modelling of health statistics*, page 187204. New York: Wiley, 2001.

S.M. LeFort, K. Gray-Donald, K.M. Rowat, and M.E. Jeans. Randomized controlled trial of a community-based psychoeducation program for the self-management of chronic pain. *Pain*, 74:297306, 1998.

D. Lindley and A. Smith. Bayes estimation for linear models. *Journal of the Royal Statistical Society, Series B*, 34:141, 1972.

R.V. Luepker, J.M. Raczynski, S. Osganian, R.J Goldberg, Jr. J.R. Finnegan, J.R. Hedges, Jr. D.C. Goff, M.S. Eisenberg, J.G. Zapka, H.A. Feldman, D.R. Labarthe, P.G. McGovern, C.E. Cornell, M.A. Proschan, and D.G. Simons-Morton. Effect of a community intervention on patient delay and emergency medical service use in acute coronary heart disease: The rapid early action for coronary treatment (react) trial. *JAMA*, 284:6067, 2000.

G. Maldonado and S. Greenland. Estimating causal effects. *International Journal of Epidemiology*, 31:422438, 2002.

C.F. Manski. Identication problems in the social sciences. in p. v. marsden (ed.). In *Sociological methodology*. San Francisco: Jossey-Banks, 1993.

W.M. Mason, G.Y. Wong, and B. Entwisle. Contextual analysis through the multilevel linear model. in s. leinhardt (ed.). In *Sociological methodology: 19831984*. San Francisco: Jossey-Bass, 1984.

J.B. McKinlay. Some contributions from the social system to gender inequalities in heart disease. *Journal of Health and Social Behavior*, 37:126, 1996.

S.M. McKinlay. The design and analysis of the observational studya review. *Journal of the American Statistical Association*, 70:503523, 1975.

A.J. McMichael. Prisoners of the proximate: Loosening the constraints on epidemiology in an age of change. *American Journal of Epidemiology*, 149:887897, 1999.

M.E.Sobel. Causal inference in the social and behavioral sciences. In *In M. E. Sobel (Ed.), Handbook of statistical modeling for the social and behavioral sciences*, page 138. New York: Plenum., 1995.

L. Meyer, N. Job-Spira, J. Bouyer, E. Bouvet, and A. Spira. Prevention of sexually transmitted diseases: A randomised community trial. *Journal of Epidemiology and Community Health*, 45:152158, 1991.

H. Morgenstern. Ecologic studies in epidemiology: Concepts, principles, and methods. *Annual Review of Public Health*, 16:6181, 1995.

J.M. Oakes. The (mis)estimation of neigborhood effects: causal inference for a practicable social epidemiology. *Social Science and Medicine*, 58:1929–1952, 2004.

T. Parsons. *The social system*. New York: Free Press, 1951.

J. Pearl. *Causality: Models, reasoning, and inference*. New York: Cambridge University Press, 2000.

V. Persky, L. Coover, E. Hernandez, A. Contreras, J. Slezak, J. Piorkowski, L. Curtis, M. Turyk, V. Ramakrishnan, and P. Scheff. Chicago community-based asthma intervention trial: Feasibility of delivering peer education in an inner-city population. *Chest*, 116:216S223S, 1999.

S.W. Raudenbush and A.S. Bryk. A hierarchical model for studying school effects. sociology of education. *Sociology of Education*, 59:1–17, 1986.

S.W. Raudenbush and R.J. Sampson. Assessing direct and indirect effects in multilevel designs with latent variables. *Sociological Methods & Research*, 28:123153, 1999.

S.W. Raudenbush and J.D. Whillms. The estimation of school effects. *Journal of Educational and Behavioral Statistics*, 20:307335, 1995.

J.M. Robins. Data, design, and background knowledge in etiologic inference. *Epidemiology*, 12:313320, 2001.

G. Rose. Sick individuals and sick populations. *International Journal of Epidemiology*, 14:3238, 1985.

S. Rose and M.J. van der Laan. Simple optimal weighting of cases and controls in case-control studies. *The International Journal of Biostatistics*, page http://www.bepress.com/ijb/vol4/iss1/19/., 2008.

S. Rose and M.J. van der Laan. Why match? investigating matched case-control study designs with causal effect estimation. *The International Journal of Biostatistics*, page http://www.bepress.com/ijb/vol5/iss1/1/., 2009.

P.R. Rosenbaum. *Observational studies*. New York: Springer, 2002.

D.B. Rubin. Practical implications of modes of statistical inference for causal effects and the critical role of the assignment mechanism. *Biometrics*, 47:12131234, 1991.

R.H. Shipley, T.D. Hartwell, W.D. Austin, A.C. Clayton, and L.C. Stanley. Community stop-smoking contests in the commit trial: Relationship of participation to costs. *Community intervention trials, Preventive Medicine*, 24:286292, 1995.

J.D. Singer. Using sas proc mixed to t multilevel models, hierarchical models, and individual growth models. *Journal of Educational and Behavioral Statistics*, 24: 323355, 1998.

H.L. Smith. Specication problems in experimental and nonexperimental social research. In *In C. C. Clogg (Ed.), Sociological methodology*, volume 20. Oxford: Basil, 1990.

P. Starr. *The social transformation of American medicine*. New York: Basic Books, 1982.

M. Susser. *Causal thinking in health sciences: Concepts and strategies of epidemiology*. New York: Oxford, 1973.

M. Susser. Should the epidemiologist be a social scientist or a molecular biologist? *International Journal of Epidemiology*, 28:10191022, 1999.

M.J. van der Laan. Estimation based on case-control designs with known prevalance probability. *The International Journal of Biostatistics*, page http://www.bepress.com/ijb/vol4/iss1/17/, 2008.

M.J. van der Laan and S. Gruber. Collaborative double robust penalized targeted maximum likelihood estimation. *The International Journal of Biostatistics*, 2010.

M.J. van der Laan and D. Rubin. Targeted maximum likelihood learning. *The International Journal of Biostatistics*, 2(1), 2006.

G. Verbeke and G. Molenbergs. *Linear mixed models in practice: A SAS oriented approach*. New York: Springer, 1997.

C. Winship and S.L. Morgan. The estimation of causal effects from observational data. *Annual Review of Sociology*, 25:659707, 1999.

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