Published in final edited form as:

J Causal Inference. 2013 May ; 1(1): 83–106. doi:10.1515/jci-2012-0011.

Estimating the Effect of a Community-Based Intervention with Two Communities

Mark J. van der Laan,

University of California – Berkeley, Berkeley, CA, USA

Maya Petersen, and

University of California – Berkeley, Berkeley, CA, USA

Wenjing Zheng

University of California – Berkeley, Berkeley, CA, USA

Mark J. van der Laan: laan@berkeley.edu; Maya Petersen: mayaliv@berkeley.edu; Wenjing Zheng: wzhen@stat.berkeley.edu

Abstract

Due to the need to evaluate the effectiveness of community-based programs in practice, there is substantial interest in methods to estimate the causal effects of community-level treatments or exposures on individual level outcomes. The challenge one is confronted with is that different communities have different environmental factors affecting the individual outcomes, and all individuals in a community share the same environment and intervention. In practice, data are often available from only a small number of communities, making it difficult if not impossible to adjust for these environmental confounders. In this paper we consider an extreme version of this dilemma, in which two communities each receives a different level of the intervention, and covariates and outcomes are measured on a random sample of independent individuals from each of the two populations; the results presented can be straightforwardly generalized to settings in which more than two communities are sampled. We address the question of what conditions are needed to estimate the causal effect of the intervention, defined in terms of an ideal experiment in which the exposed level of the intervention is assigned to both communities and individual outcomes are measured in the combined population, and then the clock is turned back and a control level of the intervention is assigned to both communities and individual outcomes are measured in the combined population. We refer to the difference in the expectation of these outcomes as the marginal (overall) treatment effect. We also discuss conditions needed for estimation of the treatment effect on the treated community. We apply a nonparametric structural equation model to define these causal effects and to establish conditions under which they are identified. These identifiability conditions provide guidance for the design of studies to investigate community level causal effects and for assessing the validity of causal interpretations when data are only available from a few communities. When the identifiability conditions fail to hold, the proposed statistical parameters still provide nonparametric treatment effect measures (albeit non-causal) whose statistical interpretations do not depend on model specifications. In addition, we study the use of a matched cohort sampling design in which the units of different communities are

matched on individual factors. Finally, we provide semiparametric efficient and doubly robust targeted MLE estimators of the community level causal effect based on i.i.d. sampling and matched cohort sampling.

Keywords

causal effect; causal effect among the treated; community-based intervention; efficient influence curve; environmental confounding

1 Introduction

There is a rich literature on estimation of individual-level causal effects using data from randomized trials and observational studies. In such studies, treatment is “assigned” to individuals in some target population, and one typically collects data on the treatment, covariates, and outcomes of a sample of individuals from this population. Causal effect estimation in this setting aims to make the best use of measured covariates to control for the fact that the treatment is empirically or theoretically a function of such covariates.

Many important causal inference problems, however, require evaluating the causal effects of treatments or exposures assigned at the community level based on data collected from samples of individuals within these communities. In particular, interest in methods for estimating the causal effects of community-based interventions has been driven by the need to evaluate the effectiveness of interventions when delivered on a large scale in realistic settings, as compared to the efficacy of the individual level counterparts of these interventions, which can frequently be evaluated in a classic randomized controlled trial. For example, understanding the impact of an early childhood nutritional intervention on children’s weight-for-age scores or subsequent educational achievement, as compared to the efficacy of nutritional supplements as delivered in the controlled context of an individual randomized trial, requires estimating the effect of the community-wide intervention. The communities of interest in such applications are typically based on a common set of characteristics or connections among their members (e.g., shared neighborhoods, schools, clinics, and work sites).

In reviewing the causal inference literature on the related topic of neighborhood effects, Oakes [1] points out the lack of rigorous discussion addressing causal inference in multilevel settings. There is now a growing body of literature addressing multilevel settings and community level effect estimation (e.g., Small et al. [2], Hanse and Bowers [3], and Imai et al. [4]), as well as discussions on causal inference in the presence of interference (which we refer to later). However, there is a paucity of literature that addresses the scenario in which data are only available from a very small number of communities (in the extreme case, one community per level of the intervention). Such a scenario arises commonly in practice, either due to logistical or financial barriers to studying a larger number of communities, or because the size of the target population of communities is fixed. (See, e.g., studies reviewed in Varnell et al. [5] and Atienza and King [6].)

Community level causal effects are commonly estimated using one of two model-based approaches: mixed models (a.k.a. conditional models) and population-average models (see, e.g., a comprehensive review of group randomized trials published in American Journal of Public Health and Preventive Medicine from 1998 through 2002 in Varnell et al. [5]). Both of these methods have limited applicability when data are only available from few communities. At least 10 communities per treatment arm are commonly recommended when generalized estimating equations are used to estimate the parameters of a population average model [7–10], while a minimum of 10–30 communities are generally recommended when fitting mixed models. Further, both mixed models and population average models typically define the effect of the intervention as a function of a parametric data-generating distribution. Background knowledge is rarely if ever sufficient to justify the parametric model employed, and misspecification can result in both biased point estimates and misleading inference.

Oakes points out that “multilevel regression models, no matter how sophisticated, appear unable to identify useful neighborhood effects from observational data,” and proposes randomized community trials as a superior alternative to observational data when estimating the effect of community level interventions. We refer the readers to Imai et al. [4] and Zhang and Small [11] for discussions on designing such experiments when resources permit. However, the success of such designs requires random allocation of the intervention among a large enough number of communities to ensure that all sources of confounding are distributed evenly across treatment arms. The benefits of randomization thus decrease with the number of communities; in the extreme case of one community per level of the intervention, randomization confers no advantage at all and the experiment is *de facto* an observational study. In this paper, we address definition, identification, and efficient estimation of the causal effect of a community level intervention when data are only available from few communities. We focus on the following data generating experiment: 1) two levels of an intervention (referred to as treatment and control, respectively) are assigned to two communities, either by the investigator or by Nature; 2) a sample of independent units (referred to here as individuals) is drawn from each of the two communities; and 3) covariates known not to have been affected by the intervention (referred to as pre-intervention covariates) and an outcome of interest are measured on each individual. The results presented can also be straightforwardly generalized to the case in which a community level intervention with two possible levels is assigned to a fixed number K communities.

We hope to contribute to the literature on community-based interventions in several ways. First, we use a rigorous causal inference framework to nonparametrically define a community level causal effect in such settings, and to establish conditions under which this effect is identified as a parameter of the observed data distribution. Specifically, the causal effect of interest is defined in terms of a hypothetical experiment where one first assigns the *treatment* level of the intervention to both communities and records individual outcomes in the combined population of the two communities, and then one *turns back the clock* to the moment before the intervention assignment, assigns the *control* level of the intervention to both communities, and records the individual outcomes of the combined population. We refer to this effect as the marginal causal effect; in addition we consider the causal effect of

the intervention in the treated community. In the case of a fixed number K communities, this definition generalizes to the causal effect of the intervention on the individual-level outcome distribution within a target population consisting of the combined population of the K communities included in the study.

Causal effects are commonly defined and identified under a “no interference” assumption [12–15]: the potential outcomes of one observation unit are assumed to be independent of the treatment received by the other units. There is a growing body of literature on causal inference in the presence of interference (e.g., [16–24]) and contagion (e.g., [25]). Related problems have also been rigorously formulated and studied in Manski [26], Graham [27], Graham et al. [28]. In this paper, we make the assumption that there is no treatment interference at either the community or individual levels.

Identification of the causal effects of interest given the two community data generating experiment described is challenging because the treated and control communities differ not only in their intervention levels, but also in a number of “environmental” factors. All individuals within a community share both the same intervention level and the same environment. Any difference between the individual outcomes in the treated and the control communities may thus be due either to an effect of the intervention or to differences in environment. In other words, the effect of the intervention is confounded by any environmental factors that differ between communities and also affect the outcome. When there are only two (or few) communities in a study, these confounders cannot be properly adjusted for, even if measured. In order to establish identifiability of causal effects in such cases, we are forced to assume that one collects pre-intervention individual covariates that are sufficient to “block” the effect of the environment on the outcome of interest (i.e., control for confounding by the environment). We formulate this assumption as an exclusion restriction assumption in a nonparametric structural equation model [29].

Particular instances of the identifiability conditions we present have been discussed in the context of group randomized trials [30]. Our goal here is to provide an explicit, general, and distribution-free statement of these conditions, with the aim of providing a common platform for discussion of causal inference in this challenging setting. We emphasize that our results are not meant to advocate or justify the use of studies with few communities. On the contrary, we provide a rigorous argument that inferring causal effects with few communities is feasible only in very limited situations, as the identifiability assumptions needed are strong. In situations where the investigator must contend with very few communities, the interpretable identifiability conditions we present may inform study design and assessment of the extent to which results can be interpreted causally.

In addition to nonparametric definition and identification of causal effects for the two community data generating experiment, we present semiparametric efficient estimators for the estimands corresponding to the marginal treatment effect and the treatment effect among the treated under these identifiability results. The targeted maximum likelihood estimators proposed employ estimators of both the distribution of the intervention given individual level covariates, and the conditional expectation of the individual level outcome given intervention level and covariates. They are double robust, in the sense that they remain

consistent if either of these nuisance parameters is estimated consistently, and they are efficient if both are estimated consistently. Further, the proposed estimators naturally integrate the state of the art in machine learning through loss-based super learning. The estimators presented have been previously described (see, e.g., [31, 32]). Here, we illustrate their utility for the estimation of a novel causal target parameter and in the context of a novel data generating experiment.

Finally, for the case of one community per treatment arm, we consider a design in which individuals are sampled from the treated and control communities by matching on a set of individual level covariates, thereby aiming to make the different communities similar in their individual covariate distributions. We provide additional conditions under which our target causal parameters will be identified given such a biased sampling scheme. This individual-level matching provides a design alternative in situations where matching at the community-level is not feasible. 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 [33], and we evaluate the theoretical gain in information for the causal effects of interest using the matched relative to independent random sampling designs.

1.1 Organization of the article

This article is organized as follows. In Sections 2–4 we address the case that one follows up a sample of individuals from each of two communities, one of which is assigned a treatment and one a control level of an intervention, and collects on each individual pre-intervention covariates and an outcome. In Section 2, we focus on identification and estimation of the marginal treatment effect. The key identifiability assumption is formulated as an exclusion restriction on an individual level nonparametric structural equation model and corresponding targeted maximum likelihood estimators (TMLE) are proposed that involve adjustment by pre-treatment covariates measured at the individual level in order to control for the confounding due to different environments. Section 3 presents an analogous identification result and TMLE for the effect among individuals in the treated community. Section 4 evaluates the benefit of matching individuals in the treated community to individuals in the control community, and proposes a TMLE for that type of design. Section 5 uses simulations to illustrate the properties of the proposed estimators for the effect among the treated, and to investigate gains in efficiency made possible through use of a matched cohort design. Section 6 summarizes how the results can be generalized to the case of K fixed communities. Additional generalizations are presented in the technical report van der Laan [34].

2 Average causal effect: two communities

We consider an experiment in which a treatment level of an intervention is assigned to the population of one community and a control level of the intervention is assigned to the population of another community. Let $A \in \{0, 1\}$ be a variable indicating the two treatment regimens ($A = 1$ for *treatment* and $A = 0$ for *control*). Let $E \in \{e_0, e_1\}$ be a variable indicating environmental factors that differ between the two communities and affect the outcome of interest, where e_0 denotes the level of these factors in the control community and

e_1 the level of these factors in the treated community. Typically, a realization of (E, A) is generated as follows: one selects two populations, whose environment defines two e -profiles, and then assigns the control level of the intervention to one of these populations, and the treatment level of the intervention to the other, giving the realized $(e_1, 1)$ and $(e_0, 0)$, respectively, for (E, A) . The combined population of the two communities represents the target population of individuals.

2.1 Observed data

Individuals are sampled from both the treatment and control populations, and pre-intervention covariates and a post-intervention outcome are measured on each sampled individual. Let W denote these individual covariates and Y denote individual outcomes in the combined population. We sample n_1 units from the treated population, providing n_1 i.i.d draws from the conditional probability distributions of (W, Y) given $(E, A) = (e_1, 1)$. Similarly, we sample n_0 units from the control population, providing n_0 i.i.d draws from the conditional distribution of (W, Y) given $(E, A) = (e_0, 0)$.

2.1.1 The causal model and target causal parameter—In this subsection we formally define the causal effect using a nonparametric structural equation model (NPSEM; [29]), and provide an explicit link between this model and the observed data. This lays the groundwork for defining our target causal effect and addressing its identifiability from the observed data.

2.1.1.1 Nonparametric structural equation model: Since our target population of units is the combined population in the given study, we assume that $E \in \{e_0, e_1\}$. We use the variable $B \equiv (E, A) \in \{(e_1, 1), (e_0, 0), (e_1, 0), (e_0, 1)\}$ to denote an individual's community and its intervention level. We assume the following causal NPSEM to describe the mechanism for assigning $(E, A) \in \{(e_1, 1), (e_1, 0), (e_0, 1), (e_0, 0)\}$ to a sampled individual from the target population, and for generating covariate and outcome data on that individual.

The NPSEM for an individual, with endogenous variables $X = (E, A, W, Y)$, is given by:

$$\begin{aligned} U_B &\sim P_{U_B} \\ B \equiv (E, A) &= f_B(U_B) \in B \equiv \{(e_1, 1), (e_0, 0), (e_1, 0), (e_0, 1)\} \\ U_{W,Y}(B) &\equiv (U_W(E), U_Y(E, A)) \sim P_U(B) \\ W &= f_W(E, U_W(E)) \\ Y &= f_Y(E, A, W, U_Y(E, A)), \end{aligned} \quad [1]$$

with $U \equiv (U_B, \{U_{W,Y}(b): b \in B\})$ denoting the exogenous variables. This encodes an experiment where an individual is drawn by first selecting which community E to draw from and what treatment A is assigned to that community; then, for this given $B = b$, one has a fixed distribution $P_U(b)$ on the errors of individual level variables; individual covariates and outcomes are drawn according to these error distributions. Note that it is assumed here that individual errors $U_{W,Y}(B)$ are i.i.d. draws from the environment–treatment-specific $P_U(B)$. Further, we assume that the individual covariates W are measured before the community intervention was implemented or, more generally, that they are not affected by the

community intervention. Therefore, W is only a function of E , and not of A . This NPSEM defines a random variable (U, X) on an individual. We denote its distribution by $P_{U,X}$.

We define the counterfactual $Y(a)$ as $Y(a) \equiv f_Y(E, A = a, W, U_Y(E, a))$, which is the outcome of an individual had the community been assigned $A = a$ while the environment E remains the same, that is, $(E, A) = (E, a)$, and the individual's covariates W are kept fixed. Note that the outcome error is now $U_Y(E, A = a)$, since the change in intervention is implemented at the community level and therefore affects the choice of environment–treatment-specific error. Similarly, we define $Y(a, w) \equiv f_Y(E, A = a, W = w, U_Y(E, a))$, $Y(e, a, w) \equiv f_Y(E = e, A = a, W = w, U_Y(e, a))$, and $W(e) \equiv f_W(E = e, U_W(e))$.

2.1.1.2 Link to observed data: Consider the study design presented above where one draws n_1 individuals from the treated community, and n_0 individuals from the control community. Thus, the observed data consists of n_1 i.i.d. draws from the $b_1 = (e_1, 1)$ -specific counterfactual $(W(b_1), Y(b_1))$ and n_0 i.i.d. draws from the $b_0 = (e_0, 0)$ -specific counterfactual $(W(b_0), Y(b_0))$. Let P_{b_0, b_1} denote the distribution of (U, X) that results from

restricting the above NPSEM by setting $B \in \{(e_1, 1), (e_0, 0)\} \sim \text{Ber}\left(\frac{n_1}{n_1 + n_0}\right)$. This way, the observed data on the $n = n_0 + n_1$ individuals can be represented as n i.i.d. copies of $O = (B, W(B), Y(B)) \sim P$ whose probability distribution P is implied by P_{b_0, b_1} . That is, one first draws

$B \in \{(e_1, 1), (e_0, 0)\}$ according to the Bernoulli distribution $\text{Ber}\left(\frac{n_1}{n_1 + n_0}\right)$; then given $B = (e, a)$, one draws $(W(B), Y(B))$ implied by the NPSEM with individual level errors drawn from $P_U(e, a)$. This representation of the two observed samples as n i.i.d. draws simplifies our presentation.

Our goal now is to define the causal effect of interest on the NPSEM, as a parameter of the distribution $P_{U,X}$, and then, under certain additional assumptions on the NPSEM, establish identifiability of this causal effect from the observed data distribution $P = P(P_{b_0, b_1})$.

2.1.1.3 Target parameter on the NPSEM: This NPSEM allows us to define the outcome distribution under set intervention, keeping the selection of the environment random, and to define corresponding causal effects of the intervention on the outcome.

We define our causal parameter of interest in the NPSEM for $X = (E, A, W, Y)$ as

$$\Psi^F(P_{U,X}) = E(Y(1) - Y(0)), \quad [2]$$

where the reader is reminded that $Y(a)$ is the counterfactual defined by setting $Y(a) \equiv f_Y(E, A = a, W, U_Y(E, a))$. This additive causal effect of A on Y corresponds to a hypothetical experiment where one first assigns treatment to both communities, records the individual outcomes, then turns back the clock and assigns control to both communities and records the individual outcomes. As is apparent from the NPSEM, this causal quantity assesses the effect of the intervention in the context of the given environments e_0 and e_1 . That is, this effect is conditional on the values of the given e_0 and e_1 , and should not be generalized to situations outside of the study without further assumptions.

2.2 Identifiability of target causal effect from observed data

We next address the identifiability of $E(Y(1) - Y(0))$ from the probability distribution $P = P(P_{b_0, b_1})$ of O . For this purpose, we make the following assumption on the NPSEM: we assume that E affects Y only via its effects on W . More specifically, we assume

$$U_Y(e_1, a) = U_Y(e_0, a) \text{ for } a \in \{0, 1\}, \text{ and } f_Y(E, W, A, U_Y(E, A)) = f_Y(W, A, U_Y(A))$$

(No direct effect of environment on individual outcome).

It follows that $Y = f_Y(W, A, U_Y(A))$. Hereafter, we also refer to this as the “no residual environmental confounding” assumption. In addition, we will assume that U_W is independent of $U_Y(a)$, $a \in \{0, 1\}$.

2.2.1 Heuristics behind “no residual environmental confounding” assumption

—The idea behind this assumption is that e_j , although common to all units in community j , results in unit specific effects of e_j on Y , which is some function $f()$ of characteristics C of the unit and e_j . 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 : for example, $W(e_1) = f(C_1, e_1)$ 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 between the two populations (i.e., $e_0 - e_1$) at the individual level.

Let us consider an example. Consider a study that is interested in evaluating the causal effect of an intervention such as a community wide program to improve early childhood nutrition, consisting, for example, in community-wide educational outreach and availability of vouchers for nutritional supplements. Consider two cohorts of children from two different regions. In one region all children are exposed to the intervention in the sense that they live in a community in which the media outreach is implemented and vouchers are available. Note that this community-level definition of exposure is chosen to reflect the policy question of interest, and differs from possible individual-level definitions of the exposure, such as use of a voucher by a child’s care-giver. The outcome measured at the individual level is weight for age at a specific time point following implementation of the intervention. A simple comparison of weight for age scores between the two cohorts is problematic since it is known that access to high quality food sources is different between the two regions (e.g., because the treatment region is located in greater proximity to a major trading center). Thus, the environments in the two regions differ in ways expected to affect the chance that a child has a low weight-for-age score: i.e, there is a higher probability of being underweight in one cohort versus the other cohort.

Which covariates should we measure to block this effect of the different region-specific access to food? Suppose we measure at the individual level the availability of food in a child’s household prior to implementation of the intervention. We include this covariate as a component of W , giving us a component of $W(e_1)$ and $W(e_0)$. One would expect that this covariate will help to block the effect of the differential access to food in the two regions on

individual child nutrition, and thereby make the “no residual environmental confounding” assumption more reasonable. In other words, one might argue that a person in the treated community and a person in the control community who have the same value for this “blocking” covariate (and for other pre-treatment covariates), would have the same probability of developing a low weight-for-age score over the course of the study, were the communities to receive the same level of the intervention.

2.2.2 Formal statement of identifiability result—We state the identifiability result formally as a theorem.

Theorem 1 NPSEM: Consider an NPSEM with structural equations for the endogenous variables $X = (E, A, W, Y)$:

$$\begin{aligned} U_B &\sim P_{U_B} \\ B \equiv (E, A) &= f_B(U_B) \in B \equiv \{(e_1, 1), (e_0, 0), (e_1, 0), (e_0, 1)\} \\ U_{W,Y}(B) &\equiv (U_W(E), U_Y(E, A)) \sim P_U(B) \\ W &= f_W(E, U_W(E)) \\ Y &= f_Y(E, A, W, U_Y(E, A)). \end{aligned}$$

and exogenous variables $U \equiv (U_B, \{U_{W,Y}(b) : b \in B\})$. Suppose the marginal distribution $\alpha \equiv P(E = e_1)$ is known.

Counterfactuals: For $a = 0, 1$, let $Y(a) \equiv f_Y(E, a, W, U_Y(E, a))$ denote the counterfactuals corresponding with setting $A = a$. We also define $(W(e), Y(e, a)) \sim P_{e,a}$ as the counterfactuals corresponding with setting $W(e) \equiv f_W(e, U_W(e))$ and $Y(e, a) \equiv f_Y(e, a, W(e), U_Y(e, a))$. Finally, we define $Y(e, a, w) \equiv f_Y(e, a, w, U_Y(e, a))$ as the counterfactual of Y corresponding with intervention $E = e, A = a, W = w$.

Observed data: Let $O = (B, W(B) \equiv W(e_B), Y(B))$, where $B \sim \text{Bernoulli}(\alpha) \in \{(e_0, 0), (e_1, 1)\}$. Conditional on $B = (e_a, a)$, the pair $(W(B), Y(B))$ is distributed as $(W(e_a), Y(e_a, a)) \sim P_{e_a}$. In particular, we note that the distribution of B in the observed data 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 . 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 $P_{e_0,0}$, and n_1 i.i.d. observations from $P_{e_1,1}$, with $n_1/(n_0 + n_1) = \alpha$.

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

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

Exclusion and Randomization assumptions on NPSEM: Suppose the following hold:

- No direct effect: E has no direct effect on Y . In other words, $U_Y(e_1, a) = U_Y(e_0, a) = U_Y(a)$ for $a \in \{0, 1\}$, and $f_Y(E, W, A, U_Y(E, A)) = f_Y(W, A, U_Y(A))$.
- Assume U_W is independent of $U_Y(a)$ for $a \in \{0, 1\}$.

Identifiability Result: Then,

$$\begin{aligned}\Psi^F(P_{U,X}) &= E_{W(B)} \{E(Y(B)|B=(e_1, 1), W) - E(Y(B)|B=(e_0, 0), W)\} \\ &\equiv \Psi(P_O) = \psi_0.\end{aligned}$$

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

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

where we used that W is independent of $U_Y(a)$, which holds since U_W, E is independent of $U_Y(a)$ and thereby $W = f_W(E, U_W)$ is independent of $U_Y(a)$. We note that $P(W=w) = P(W(e_1) = w | E = e_1)P(E = e_1) + P(W(e_0) = w | E = e_0)P(E = e_0) = P_{W(e_1)}(w)\alpha + P_{W(e_0)}(w)(1 - \alpha)$.

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

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

where we use that U_W is independent of $U_Y(a)$.

Similarly, $E(Y(B)|W(B)=w, B=(e_0, 0)) = Ef_Y(0, w, U_Y(0))$. 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, this proves

$$\psi_0 = \sum_w \{Ef_Y(A=1, w, U_Y(1)) - Ef_Y(A=0, w, U_Y(0))\} P(W=w),$$

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

2.2.3 Commitment to statistical parameter and model for observed data—The model for the probability distribution P_O of $O = (B, W(B), Y(B))$ implied by the NPSEM is nonparametric. Based on theorem 1, we propose the following target parameter of the

distribution of the observed data structure $O = (B, W(B), Y(B))$ with B Bernoulli in $\{(e_1, 1), (e_0, 0)\}$:

$$\Psi(P_O) \equiv E_{W(B)}\{E(Y(B)|B=(e_1, 1), W) - E(Y(B)|B=(e_0, 0), W)\}. \quad [3]$$

This statistical parameter corresponds to a statistical experiment where, for each stratum of the individual covariate in the combined population, one obtains the difference in expected outcome for the treated community and the expected outcome for the control community, then averages this difference over the covariate distribution in the combined population. At its face value, this statistical parameter provides a nonparametric treatment effect measure. The causal interpretation of this treatment effect measure, however, is contingent on the non-testable assumptions A1–A2. In particular, with respect to A1, we note that even if W does not succeed in capturing the complete effect of e on the unit-specific outcome Y , (in other words, one is not able to establish identifiability of the target causal parameter), adjusting for a rich set of covariates W 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 . Thus, the statistical estimand eq. [3], fully adjusted for W , may still be of interest as a closest possible approximation to the wished for causal effect. Of course, in such a case, the resulting estimate should be carefully and transparently interpreted as at best an approximation.

2.3 Estimation and inference

The targeted maximum likelihood estimator of the statistical parameter eq. [3] has been defined previously and statistical inference for this estimator has been described (see, e.g., van der Laan and Rubin [35], van der Laan and Rose [36] for the targeted MLE, and van der Laan and Gruber [37], Gruber and van der Laan [38]) for the collaborative targeted MLE).

Since we have arrived at the statistical estimation stage, we will denote $O = (B, W = W(B), Y = Y(B))$ and the pooled sample with $O_i, i = 1, \dots, n = n_1 + n_2$. We will also use $B = a$ to denote the realization $B = (E = e_a, A = a)$, for $a = 0, 1$. Note that the data generating distribution P_O is determined by the marginal distribution P_W of W , the conditional distribution $P_{B|W}$ of B , given W , and the conditional distribution $P_{Y|B, W}$ of Y , given B, W . The parameter $\Psi(P_O)$ depends on P_O through both P_W and $Q_{Y,0}(B, W) \equiv E_0(Y|B, W)$, as well as the “treatment” mechanism $P_{B|W}$. We will denote the treatment mechanism with g_0 and the other two factors of the likelihood with Q_0 .

Implementation of the estimator begins with application of a data adaptive algorithm such as super learner (van der Laan et al. [39], Polley and van der Laan [36], van der Laan and Rose

[36]) to fit $\bar{Q}_{Y,0} = E_0(Y|B, W)$. This initial estimate $\bar{Q}_{Y,n}^0$ is subsequently updated using targeted maximum likelihood estimation. The marginal distribution of W is estimated with the empirical of the pooled sample $W_i, i = 1, \dots, n = n_1 + n_2$. The targeted maximum likelihood estimate also requires a fit of $P(B = 1 | W)$. The estimator is double robust in the sense that it remains unbiased if one either consistently estimates $g_0(1 | W) \equiv P(B = 1 | W)$ or $\bar{Q}_{Y,0}(B, W) \equiv E(Y|B, W)$. The estimator is efficient if the initial estimator $\bar{Q}_{Y,n}^0$ is

consistent and g_n is consistent as well. If g_n is misspecified but $\bar{Q}_{Y,n}^0$ is consistent, the estimator can be either super efficient or inefficient, depending on the limit of g_n . The targeted maximum likelihood estimator can be further refined with the collaborative targeted maximum likelihood estimation method, resulting in a collaborative double robust estimator which has generally better finite sample efficiency and is consistent under weaker conditions.

The double robustness of the targeted maximum likelihood estimator in terms of the components $g_0, \bar{Q}_{Y,0}$ of the distribution of observed data structure (B, W, Y) , translates into the following robustness in terms of the underlying counterfactual distributions $P_{e_0,0}$ and $P_{e_1,1}$ of the two samples. Firstly, we note that $P(w) \equiv P(W = w) = \alpha P_{W(e_1)}(w) + (1 - \alpha)P_{W(e_0)}(w)$. We also define $Q_{1,X}(w) = E(Y(e_1, 1)|W(e_1) = w)$, $Q_{0,X}(w) = E(Y(e_0, 0)|W(e_0) = w)$, and we note that

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

Thus, the double robustness of the targeted MLE for estimation of ψ_0 in a nonparametric model for $O = (B, W, Y)$ in terms of $g_0, \bar{Q}_{Y,0}$ can be restated as follows: the targeted MLE will be consistent if either the outcome regressions $Q_{1,X}, Q_{0,X}$ on the covariates are consistently estimated for both samples, or if the ratio $P_{W(e_1)}/P_{W(e_0)}$ of the covariate distributions for the two samples is consistently estimated. In particular, the identifiability condition $0 < P(B = 1|W) < 1$ almost everywhere (a.e.) translates into $0 < \alpha < 1$, and that the Radon–Nykodym derivatives $P_{W(e_0)}(w)/P_{W(e_1)}(w) < \infty$ and $P_{W(e_1)}(w)/P_{W(e_0)}(w) < \infty$ for the covariate distributions are bounded. Thus, if a covariate can have a certain value in the treated population, then that value should also occur in the control population, and vice versa.

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

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

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 $g_n, \bar{Q}_n$ for $g_0, \bar{Q}_{Y,0}$.

3 Causal effect among the treated: two communities

Using the same NPSEM and notations as in theorem 1, we now consider an alternative target causal parameter, referred to as the treatment effect among the treated.

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

This causal parameter corresponds to an effect assessed by an ideal experiment where after recording the individual outcomes of the treated community, the investigator can turn back the clock and implement the *control* intervention on this same community, and record its individual outcomes under the control regimen.

A similar argument as in theorem 1 shows that if the following assumptions on the NPSEM hold:

- No direct effect: E has no direct effect on Y , that is, $f_Y(E, W, A, U_Y(E, A)) = f_Y(W, A, U_Y(A))$.
- Randomization: U_W is independent of $U_Y(a)$, $a \in \{0, 1\}$,

then

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

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

We consider estimation and inference for the alternative target statistical parameter implied by this result:

$$\psi_0 = \Psi(P_0) = E_{W(e_1)} \{E(Y | B = (e_1, 1), W) - E(Y | B = (e_0, 0), W)\}, \quad [4]$$

Note that eq. [4] differs from the statistical parameter eq. [3] in that the outcome differences are averaged over the covariate distribution of the treated community alone.

As we discuss below, the identifiability of eq. [4] has weaker requirements for data support than that of eq. [3]. More specifically, one only needs that $P_{W(e_1)}/P_{W(e_0)} < \infty$, that is, the support of $P_{W(e_0)}$ contains the support of $P_{W(e_1)}$. Moreover, in the following sections, we show that an individually matched cohort design is particularly optimal for targeting this effect among the treated.

3.1 Targeted MLE for the causal effect among the treated

In this subsection we consider estimation of the statistical parameter eq. [4] defined above. Similar to the previous section, we denote the observed data as $n = n_1 + n_2$ i.i.d. observations

of $O = (B, W(B), Y(B))$, where B is binary. We make no assumptions on the data generating distribution P_O ; that is, we assume a nonparametric statistical model.

3.1.1 The efficient influence curve of the target parameter—Firstly, consider the parameter $P \mapsto \Psi_b(P) \equiv E_P(E_P(Y | B = b, W) | B = 1)$. The efficient influence curve of this parameter is presented in van der Laan and Rose [36] and technical report van der Laan [34], and is given by

$$D_b^*(P) = \frac{I(B=b)}{P(B=1)} \frac{g(1|W)}{g(b|w)} (Y - \bar{Q}_Y(b, W)) + \frac{I(B=1)}{P(B=1)} (\bar{Q}_Y(b, W) - \Psi_b(P)).$$

Here $\bar{Q}_Y(P)(b, W) = E_P(Y | B = b, W)$ and $g(P)(b|W) = P(B = b | W)$.

Thus, the efficient influence curve of $\Psi(P) = \Psi_1(P) - \Psi_0(P)$ is given by

$$D^*(P) = \left\{ \frac{I(B=1)}{P(B=1)} - \frac{I(B=0)}{P(B=1)} \frac{g(1|W)}{g(0|W)} \right\} (Y - \bar{Q}_Y(B, W)) + \frac{I(B=1)}{P(B=1)} (\bar{Q}_Y(1, W) - \bar{Q}_Y(0, W) - \Psi(P)).$$

3.1.2 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 $P(B = 1)$. We note that this estimating function is double robust in the sense that it is an unbiased estimating function for $\psi_0 \equiv \Psi(P_O)$, if either Q is correctly specified or g is correctly specified. Formally, this is stated as

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

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

One can 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 .

3.1.3 Targeted maximum likelihood estimator—We next explain the targeted maximum likelihood algorithm that maps an initial estimator $\hat{P}$ into a targeted fit $\hat{P}^*$. Suppose Y is bounded in the unit interval. Given an initial estimator $\bar{Q}_{Y,n}$ of $E(Y|B, W)$, an initial estimator g_n of $P(B|W)$, and the empirical distribution of W , in order to compute the

targeted MLE we define the fluctuation $\text{logit}Q_{Y,n}(\varepsilon_1)(B, W) \equiv \text{logit}Q_{Y,n}(B, W) + \varepsilon_1 C_1(g_n)(B, W)$, and $\text{logit}(g_n(\varepsilon_2)(1|W)) \equiv \text{logit}(g_n(1|W)) + \varepsilon_2 C_2(P)(W)$, where

$$C_1(g) \equiv \frac{I(B=1)}{P(B=1)} - \frac{I(B=0)}{P(B=1)} \frac{g(1|W)}{g(0|W)},$$

$$C_2(P) \equiv \frac{1}{P(B=1)} \{Q(P)(1, W) - Q(P)(0, W) - \Psi(P)\}.$$

These two one-dimensional fluctuations of the regression $Q_{Y,n}$ and the treatment mechanism g_n represent a fluctuation $P(\varepsilon)$ of P , where the empirical distribution of W is held 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 $Q_{Y,n}$ as an off-set, and ε_2 is estimated with standard linear logistic regression fixing $g_n(1|W)$ as offset in the logistic regression model for $P(B = 1|W)$.

This maximum likelihood estimator $\varepsilon_n = (\varepsilon_{1n}, \varepsilon_{2n})$ now defines an update $P^1 = P(\varepsilon_n)$. This targeted maximum likelihood updating is iterated till convergence and the final P^* ,

identified by a $\bar{Q}_{Y,n}^*$, g_n^* (and the empirical for P_W), is called the targeted maximum likelihood estimator of the distribution P_O , while $\Psi(P^*)$ is called the targeted maximum likelihood estimator of ψ_0 . The targeted maximum likelihood estimator $\Psi(P^*)$ solves the efficient influence curve estimating equation,

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

Note that we can use machine learning/super learning to obtain the initial P (i.e., $Q_{Y,n}$ and g_n).

Since P^* solves, in particular,

$$0 = \frac{1}{n} \sum_{i=1}^n \frac{I(B_i=1)}{P(B=1)} \left\{ \bar{Q}_{Y,n}^*(1, W_i) - \bar{Q}_{Y,n}^*(0, W_i) - \Psi(\hat{P}^*) \right\},$$

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

$$\Psi(\hat{P}^*) = \frac{1}{n} \sum_{i=1}^n \frac{I(B_i=1)}{P(B=1)} \left(\bar{Q}_{Y,n}^*(1, W_i) - \bar{Q}_{Y,n}^*(0, W_i) \right),$$

that is, as the empirical mean of $\bar{Q}_{Y,n}^*(1, W) - \bar{Q}_{Y,n}^*(0, W)$ among the observations with $B_i =$

1. This evaluation makes use of g_n^* only through $\bar{Q}_{Y,n}^*$.

4 Matching the treated and control community at individual level

We now understand the estimator of the statistical parameter ψ_0 and the conditions under which this statistical parameter ψ_0 equals the target additive causal effect ψ_0^F (marginally or among the treated). From this, we conclude the importance of measuring individual characteristics that can “block” the effect of the two communities’ differing environments (i.e., e_0 and e_1) on the individual outcome, so that units in the treated population with $W = w$ are exchangeable with units in the control population with $W = w$ with respect to their 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 also be very different, thereby possibly generating lack of experimentation for g_0 : i.e., $g_0(1|W)$ may approach 0 or 1 for some W -values. Even if such imbalances do not result in non-identifiability, large imbalances can still increase the asymptotic variance of the targeted MLE; 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} become 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 crucial that the two populations are fairly comparable with respect to the factors e that have an impact on the outcome. The more comparable the communities, the smaller the asymptotic variance of the targeted MLE adjusting for W will be.

Despite the desirability of comparable treatment and intervention communities, in many cases achieving this comparability is not feasible. In particular, the investigator may have minimal control over selection of the study communities. Nonetheless, in such settings, the investigator may maintain some control over how individuals from the study communities are sampled. This suggests the potential utility of an individually matched cohort design, in which a unit from the treated population is matched with a unit from the control population based on a set of individual level variables that are not affected by the treatment.

4.1 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 combined population of the two communities, and representing the two sample problem of sampling n_0 observations from $P_{U(e_0, 0)}$ and n_1 observations from $P_{U(e_1, 1)}$. 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 M for P_O is nonparametric.

As above, we consider the following two target statistical parameters of P_O :

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

We saw previously that under the exclusion and randomization assumptions A1 and A2, ψ_0 equals $E(Y(1) - Y(0))$, the marginal causal effect of the intervention on the individuals in the combined population of the two communities, and ψ_0^t equals $E(Y(1) - Y(0)|E = e_1)$, the causal effect on the individuals of the treated community. 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.

Recall that the efficient influence curves for ψ_0 and ψ_0^t are given by

$$D^*(Q_0, g_0, \psi_0)(O) = \left\{ \frac{I(B=1)}{g_0(1|W)} - \frac{I(B=0)}{g_0(0|W)} \right\} (Y - \bar{Q}_{Y,0}(B, W)) + \bar{Q}_{Y,0}(1, W) - \bar{Q}_{Y,0}(0, W) - \psi_0$$

and

$$D^{*t}(Q_0, g_0, \psi_0^t) = \left\{ \frac{I(B=1)}{P(B=1)} - \frac{I(B=0)}{P(B=1)} \frac{g_0(1|W)}{g_0(0|W)} \right\} (Y - \bar{Q}_{Y,0}(B, W)) + \frac{I(B=1)}{P(B=1)} (\bar{Q}_{Y,0}(1, W) - \bar{Q}_{Y,0}(0, W) - \psi_0^t),$$

respectively.

4.2 Matched cohort sampling

As an alternative to 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 consider the following J -to-1 matched cohort sampling:

1. Let $M \subset W$ be a subset of the individual level covariates W that we will match on.
2. Sample $(W(e_1), Y(e_1, 1))$ from the conditional distribution of (W, Y) , given $B = (e_1, 1)$. Let m_1 denote the observed value of $M(e_1)$.
3. Sample J times $(W(e_0), Y(e_0, 0))$ from the conditional distribution (W, Y) given $B = (e_0, 0)$ and $M(e_0) = m_1$.
4. Let

$$O^m \equiv ((W(e_1), Y(e_1, 1)), (W(e_0)^j, Y(e_0, 0)^j), j=1, \dots, J),$$

be the cluster of matched observations.

5. Repeat this experiment n times, resulting in n clusters $O_i^m, i = 1, \dots, n$.
6. Note 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 independent copies.

4.3 Estimation in matched cohort designs

Matched cohort designs provide a biased sample from the distribution, P_o , so that a new identifiability result is required since the previous identifiability results were based on sampling i.i.d. copies of $O = (B, W, Y)$. Targeted maximum likelihood estimation, and efficient estimation in general, based on this type of biased sampling, including matched

case-control/cohort sampling, was studied in van der Laan [33] and Rose and van der Laan [39]. This previous work assumed that the following quantities are known

$$q_0 = P(B=1) = \frac{n_1}{n_0 + n_1}$$

$$\bar{q}_0(M) = \frac{q_0}{P(B=1|M)} P(B=0|M).$$

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 $O = (B, W, Y)$, but giving each observation a weight. The observations with $B_i = 1$ are assigned weights $q_0 = n_1/(n_0 + n_1)$. The J observations with $B_i = 0$, which were matched to a $B_i = 1$, receive weights $q_0(M_i)/J$. The resulting case-control weighted targeted MLE now targets the same parameter ψ_0 or ψ_0^t , is asymptotically efficient, and has the same double robustness property as the targeted MLE applied to the i.i.d. sampling of $O = (B, W, Y)$. The efficient influence curve of ψ_0, ψ_0^t for this matched cohort sampling model, as given below, can as usual be used for statistical inference based on the case-control weighted targeted MLE.

4.3.1 Knowledge needed to determine weights used to correct for matched sampling

—Suppose that 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 . Of course, the required weights $q_0(M)$ are now determined by the Bayes rule.

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

In van der Laan [33] 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,

$$D^m(Q, g, \psi_0)(O^m) = q_0 D^*(Q, g, \psi_0)(B=1, W(e_1), Y(e_1, 1)) + \frac{\bar{q}_0(M(e_1))}{J} \sum_{j=1}^J D^*(Q, g, \psi_0)(B=0, W(e_0^j), Y(0, e_0^j))$$

for the marginal causal effect ψ_0 , and

$$D^{mt}(Q, g, \psi_0^t) = q_0 D^{*t}(Q, g, \psi_0^t)(B=1, W(e_1), Y(e_1, 1)) + \frac{\bar{q}_0(M(e_1))}{J} \sum_{j=1}^J D^{*t}(Q, g, \psi_0^t)(B=0, W(e_0^j), Y(0, e_0^j))$$

for the causal effect among the treated ψ_0^t . This design includes the “no-matching” choice by setting M equal to empty set, and $J = n_0/n_1$, in which case $q_0(M) = 1 - q_0$, 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 i.i.d. samples study design.

4.4.1 Evaluating the benefit of matching in the design—The variance of the efficient influence curve for a parameter 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^{mt} we can investigate if matching decreases 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. It follows from the formulas above that the efficient influence curve for ψ_0 based on sampling the cluster observation O^m is given by

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

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

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

Firstly, it is good to see that indeed, if M is empty, and thereby $q_0(M) = 1 - q_0$, then D^m and D^{mt} correspond with i.i.d sampling of $O = (B, W, Y)$ and the respective efficient influence curves D^* and D^{*t} , for ψ_0 and ψ_0^t given above. In other words, D^m and D^{mt} simply combine observations from control and treatment sample in the cluster O^m , but if the observations in a cluster are independent, this coupling serves no purpose beyond that it allows us to compare the efficient influence curve under matching with the efficient influence curve under the original no-matching two sample design. Therefore, the question “Is matching

improving the design with respect to estimation of the target parameter?” corresponds with “Do the variances of the efficient influence curves D^m and D^{mt} become smaller as the subset M becomes closer to W ?”

To answer this question we simply write down the efficient influence curves for the cases of $M = \emptyset$, $M \subset W$ and $M = W$, for both target parameters. With a slight abuse to notation, we write $g_0(1) \equiv q_0$ and $g_0(1|M) \equiv P(B = 1|M)$ to stress that these are related to the conditional probability distribution $g_0(\cdot|W) = P_O(B = \cdot|W)$ of B , given W . This means that we can denote $g_0(M) = \{g_0(1)/g_0(1|M)\}g_0(0|M)$.

Marginal Causal effect, no matching ($M = \emptyset$)

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

Marginal Causal effect, partial matching ($M \subset W$)

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

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

$$\begin{aligned} D^m = & \frac{g_0(1)}{g_0(1|W(e_1))} \{Y(e_1, 1) - \bar{Q}_{Y,0}(1, W(e_1))\} \\ & + g_0(1) \{ \bar{Q}_{Y,0}(1, W(e_1)) - \bar{Q}_{Y,0}(0, W(e_1)) - \Psi(Q_0) \} \\ & - \frac{g_0(1)}{g_0(1|W(e_0))} \{Y(e_0, 0) - \bar{Q}_{Y,0}(0, W(e_0, 0))\} \\ & + g_0(0|W(e_0)) \{ \bar{Q}_{Y,0}(1, W(e_0)) - \bar{Q}_{Y,0}(0, W(e_0)) - \Psi(Q_0) \}, \end{aligned}$$

where $W(e_0) = M(e_1) = W(e_1)$ with probability 1. We note that the inverse weighting by $g(0|W)$ and $g(1|W)$ 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 ($M = \emptyset$)

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

Causal effect among treated, partial matching ($M \subset W$)

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

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

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

Note that the estimator ψ_n^t that solves the efficient influence curve equation $P_n D^{mt}(Q_n, g_n, \psi) = 0$ is given by $\psi_n^t = P_n Y(e_1, 1) - P_n Y(e_0, 0)$, a difference of sample means between the two groups, where the observations $Y(e_1, 1)$, $Y(e_0, 0)$ 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 .

4.4.2 Remark: matched case-control sampling vs. matched cohort sampling—

It was shown in Rose and van der Laan [40], by practical demonstration in simulation studies, that matched case-control sampling designs typically provide less information about the parameter of interest than non-matched case-control designs. On the other hand, we see from the above formulas that matched cohort sampling is likely to improve upon unmatched cohort sampling. This distinction can be understood by noting that in order for matching to be effective, the weights $q_0(M_1)$ ought to cancel/stabilize the inverse treatment weightings in the i.i.d.-based efficient influence curve. In matched cohort sampling, we see that this is indeed the case since one conditions on the “treatment” B . However, in matched case-

control samplings, one condition on the outcome, therefore, the weights $q_0(M_1)$ introduce additional inverse outcome weighting, without stabilizing the treatment weights.

4.4.3 Summary: optimizing the design—If one can select two populations for which the pre-treatment covariate distributions P_{e1}, P_{e0} are almost equivalent, that is, $P_{e1}(W) \approx P_{e0}(W)$, while W blocks any effect of E on Y , then that implies that $P(B = 1|W) \approx P(B = 1)$, 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 .

5 Simulations

In previous sections we showed that under the exclusion and randomization assumptions A1 and A2, causal effects of the intervention are identifiable from the observed data of the combined population. Moreover, the resulting statistical parameters are equivalent to those arising from a causal model in which the intervention is assigned at the individual level. Properties of the targeted MLE for this estimand, which corresponds to the marginal treatment effect under identifiability results presented here, have been illustrated in previous work (e.g., [41]). In this section, we illustrate the double robustness of the targeted MLE for the estimand corresponding to the causal effect among the treated, under both the independent sampling and matched cohort sampling two community designs. In addition, we use simulations to investigate cases in which the matched cohort design provides an efficiency gain over an independent sampling design when using a substitution estimator.

5.1 Data generation

We evaluated the relative efficiency of the matched cohort design, compared to independent sampling, while varying the extent to which a matching variable W_2 affects the outcome Y and the extent to which it predicts the intervention A (or equivalently, community membership). We considered 3 levels (low, medium, high) for each of these parameters, resulting in $3 \times 3 = 9$ data generating distributions. The data generating process for an individual i is described below.

$$\begin{aligned} E &\sim U(0, 2) \\ W_{1,i}|E &\sim 10 \times \text{Beta}(\alpha=1.2, \beta=1+1.1E) \\ W_{2,i}|E &\sim P_{W_{2,k}} \\ Y_i|W_i, A &= \bar{Q}_{0,i} + N(0, 3^2), \end{aligned}$$

where $P_{W_{2,k}}$ ranges over three distributions with different effects of E on W_2 :

$$\begin{aligned}
P_{W_2,low} &= Multinom(p_1 = \Phi(-0.05 + 0.01E), \\
p_2 &= \Phi(-0.08 + 0.01E) (1 - \Phi(-0.05 + 0.01E)), \\
p_3 &= (1 - \Phi(-0.08 + 0.01E)) (1 - \Phi(-0.05 + 0.01E))) \\
P_{W_2,medium} &= Multinom(p_1 = \Phi(-0.18 + 0.3E), \\
p_2 &= \Phi(-0.2 + 0.05E) (1 - \Phi(-0.18 + 0.3E)), \\
p_3 &= (1 - \Phi(-0.2 + 0.05E)) (1 - \Phi(-0.18 + 0.3E))) \\
P_{W_2,high} &= Multinom(p_1 = \Phi(-1.2 + 0.6E), \\
p_2 &= \Phi(-1 + 0.8E) (1 - \Phi(-1.2 + 0.6E)), \\
p_3 &= (1 - \Phi(-1 + 0.8E)) (1 - \Phi(-1.2 + 0.6E)));
\end{aligned}$$

and the outcome conditional mean $\bar{Q}_{0,l}$ ranges over three functions with different effects of W_2 on Y :

$$\begin{aligned}
\bar{Q}_{0,low} &= 1 + 2.1A - 0.1W_{2,i}^2 + 1.5W_{1,i}^2 + A \times (3W_{1,i} + 0.1W_{2,i}^2) \\
\bar{Q}_{0,medium} &= 1 + 2.1A - 1.5W_{2,i}^2 + 1.5W_{1,i}^2 + A \times (3W_{1,i} + 1W_{2,i}^2) \\
\bar{Q}_{0,high} &= 1 + 2.1A - 5W_{2,i}^2 + 1.5W_{1,i}^2 + A \times (3W_{1,i} + 2W_{2,i}^2).
\end{aligned}$$

The effect size of E on W_2 determines the extent to which $g_0(A = 1|W_2)$ depends on W_2 . Under $P_{W_2,low}$, $g_0(A = 1|W_2)$ takes values roughly 0.5 for all values of W_2 ; under $P_{W_2,medium}$, it takes values 0.6, 0.57, 0.52 for $W_2 = 1, 2, 3$, respectively; under $P_{W_2,high}$, it takes values 0.8, 0.77, 0.42.

As mentioned previously, when one only has two communities, the causal effect assessed is in terms of the two observed communities. Therefore, we fixed the observed communities to be $E \in \{0.001, 1.999\}$, and assigned $A = 1$ to community $e = 1.999$ using a single Bernoulli trial. The outcome of this trial determined the distributions of individual covariates and outcome in each community, as well as the value of the true causal effect parameter. All samples were drawn from these distributions.

5.2 Implementation of sampling schemes

Independent sampling was implemented as n_j independent draws with replacement from each community j . We used a balanced design with $n_1 = n_2$, so that there were $2 \times n_1$ independent observations under the independent sampling design.

Matched cohort sampling with one-to-one matching was implemented as follows: we drew an independent sample of size n_1 from the treated community; for each observation i in this sample, with $W_2 = w_{2,i}$, we then drew an individual from the control population by drawing $W_{1,i}$ conditional on $E = e_0$ (call its realization $w_{1,i}$) and Y_i conditional on $((w_{1,i}, w_{2,i}), A = 0)$. The true probability distribution $g_0(B|W_2)$ was used in computing the weights for each matched control. There were n_1 independent pairs of matched observations under this matched design.

5.3 Estimators

The following estimators were applied to each dataset. Note that TMLE was implemented with the logistic fluctuation, with the proper linear transformations.

- $\Psi(Q_{unadj})$: Substitution estimator using the estimator $\hat{Q}_{unadj}$ of Q_0 which only adjusts for A but not individual level covariates.
- $\Psi(\hat{Q}_{unadj}^*(g_c))$: TMLE with initial estimator $\hat{Q}_{unadj}$ and a correctly specified $g(B|W)$, which is obtained using the correctly specified $P(W|E)$ and Bayes rules.
- $\Psi(\hat{Q}_m)$: Substitution estimator using a misspecified estimator of Q_0 : $(Y \sim A + W_1 + W_2)$
- $\Psi(\hat{Q}_m^*(g_c))$: TMLE with initial estimator $\hat{Q}_m$ and a correctly specified g .
- $\Psi(\hat{Q}_c)$: Substitution estimator using the correctly specified estimator $\hat{Q}_c$ of Q_0 .
- $\Psi(\hat{Q}_c^*(g_m))$: TMLE with initial estimator $\hat{Q}_c$ and a misspecified treatment mechanism $g_m(1|w) = .5, \forall w$.
- $\Psi(\hat{Q}_c^*(g_c))$: TMLE with initial estimator $\hat{Q}_c$ and a correctly specified g .

5.4 Results: double robustness of TMLE

We illustrate the double robustness of the TMLE for the matched and independent designs, respectively, using the data generating distribution with $P_{W_2} = P_{W_2, high}$ and $E(Y|A, W) = Q_{0, high}$. The value of the causal effect among the treated in these simulations is 23.48. For each sampling design and for two choices of n_j (100 and 1,000), bias, variance, and mean squared error (MSE) were estimated over 1,000 datasets. The results in Tables 1 and 2 demonstrate that when g was correctly specified, the targeting step reduced bias in a misspecified initial outcome model. Further, the mean squared error of the unbiased estimators decreased at a rate proportional to n .

5.5 Results: when does the matched cohort design provide efficiency gain?

We investigated the relative efficiency of the matched cohort design as compared to the independent design using targeted and non-targeted substitution estimators. For sample sizes of $n_1 = n_2 = 5000$ and for each data generating distribution, we compared the ratio of the sample variances of each estimator over 1,000 datasets under independent versus matched sampling.

Results of these simulations (Table 3) show that in general there was no loss of efficiency as long as data from the matched design were analyzed with the targeted estimator. The matched design, when analyzed with the targeted estimator, provided more meaningful gains in efficiency when the matching variable strongly predicted the intervention. When the initial outcome model ($\hat{Q}_{unadj}$) did not adjust for the matching variable or other confounders, then matching and provided a great efficiency gain when analyzed with the targeted estimator, while matching without targeting did not improve efficiency. Adjustment for confounders, even using a misspecified initial outcome model, reduced the relative gain in

efficiency provided by matching and targeting. When the initial outcome model was correctly specified (Q_c), the efficiency gain provided by matching was further reduced, and the relative efficiency gains were equivalent when data were analyzed with the targeted or non-targeted estimator.

5.6 Results: variance of D^* in matched vs independent sampling

Recall that $\text{var}(D^*)$ is in fact the Cramer–Rao lower bound for the variance of a locally unbiased and asymptotically linear estimator of the parameter in a nonparametric model. On the other hand, an estimator that is asymptotically linear with influence curve D will have asymptotic variance $\text{var}(D)$. Therefore, we can assess the theoretical gain in efficiency of using matching and targeting by comparing the variance of $D^*(O)$ vs variance of $D^{m,*}(O^m)$, at $g = g_0$, and $Q = Q_0$, Q_m , or Q_{unadj} . To this end, for each data generating distribution and sampling design, we used a sample of $n_1 = n_2 = 100,000$ observations. The limits Q_m and Q_{unadj} were approximated using the sample. $\text{Var}(D^*(O))$ was approximated with the sample variance $\hat{Qar}(D^*(O))$ of $D^*(O)$ over the $2 \times 100,000$ independent observations, while $\text{Var}(D^{m,*}(O^m))$ was approximated with the sample variance $\hat{Qar}(D^{m,*}(O^m))$ of $D^{m,*}(O^m)$ over 100,000 independent pairs of matched observations.

Tables 4–6 show that when the matching variable W_2 was a strong predictor of A , the matched cohort design provided a significant gain in efficiency over the independent design, a gain which was more pronounced when the outcome model was incorrect (Q_m and $Q_{unadjust}$). On the other hand, the effect size of the matching variable on outcome seemed to have little bearing on efficiency.

6 Generalization to K communities

Full generalization of the results presented here to K fixed communities is presented in the technical report van der Laan [34]. Let $\{e_1, \dots, e_K\}$ denote the set of K communities. An NPSEM analogous to eq. [1] is defined by using the set $B = \{(e_k, a): k = 1, \dots, K, a \in A\}$. Under this NPSEM, the counterfactual $Y(a)$ and causal parameter $EY(a)$ are defined as in Section 2.1.

In a real experiment, one only observes environment–treatment combinations $\{(e_k, a_k): k = 1, \dots, K\}$. Therefore, the observed data may be considered as sampling from a restricted NPSEM, where B is restricted to $B = \{(e_k, a_k): k = 1, \dots, K\}$. More specifically, our observed data $O = (B, W, Y)$ correspond with observing K samples of n_k i.i.d. observations from $P_U(e_k, a_k)$, $k = 1, \dots, K$. The identifiability result stated in the theorem in van der Laan [34] teaches us that, under the same exclusion restriction and randomization assumption, the causal parameter $EY(a)$ corresponds with the statistical parameter

$$EY(a) = \Psi_a(P_O) = E_{W(B)} E(Y(B) | B \in B_a, W),$$

where $B_a = \{(e_k, a_k): a_k = a\}$ represents the communities for which the community-based intervention was set at level a . For example, if the intervention is binary, then we can

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

$$\Psi(P_O) = E_{W(B)} \{E(Y(B)|B \in B_1, W) - E(Y(B)|B \in B_0, W)\}.$$

The efficient influence curve, targeted MLE, collaborative targeted MLE, and statistical inference based on an estimate of the efficient influence curve correspond exactly with those presented earlier for the statistical target parameter $E_W\{E(Y|A = 1, W) - E(Y|A = 0, W)\}$ based on observing n i.i.d. copies of (A, W, Y) with $Y = Y(B)$, $W = W(B)$, and $A = I(B \in B_1)$. Thus, the practical conclusion is that one can create a single combined sample from the K community-specific samples, reduce each observation to (A, W, Y) by ignoring the data on environmental factors, and apply the targeted MLE for $E_W\{E(Y|A = 1, W) - E(Y|A = 0, W)\}$.

7 Summary and concluding remarks

This article provides contributions to the literature on causal inference for community-based interventions, concerning inference conditional on a given finite set of communities and asymptotics in the number of individuals sampled, under the assumption of no-interference.

For the experiment in which we observe two populations under two different treatment regimens, while collecting data at the individual level, the article makes the following contributions. We define the causal effect of a treatment versus control level of an intervention as the difference in the mean individual outcome under an experiment where one assigns the treatment level to both communities versus under an experiment where one assigns the control level to both communities. We establish that this causal effect can be identified under a randomization assumption, together with an exclusion restriction that assumes that measured individual level covariates which are not affected by the treatment are sufficient to block the effect of any environmental factors that differ between the two communities on the outcome. Covariates in this setting, while remaining crucial for control of confounding, thus play a slightly different role than in classical causal inference for individual-based treatments. The main purpose of the covariates here 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, repeating the process 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.

We emphasize that the identifiability conditions required are both strong and untestable – they concern the causal mechanism, and are not related to any statistical model or analytical methods used by the researcher. The strength of the assumptions required should serve as a warning that inferring causal effects of community-based interventions when data are available from only a few communities is sensible only in very limited cases, and with a very careful study design. The statistical parameters eqs. [3] and [4] provide nonparametric measures for quantifying the intervention effect. However, when the identifiability

conditions are not satisfied, the investigator is warned against moving beyond a statistical to a casual interpretation of these parameters. Nonetheless, by providing a nonparametric measure of treatment effect, these statistical parameters are more interpretable than those defined based on parametric model specifications.

We present (collaborative) targeted maximum likelihood estimators for the statistical parameters provided by these identifiability results. The estimators are based on first using super learning to estimate the outcome regression, and subsequently targeting this initial fit using a targeted maximum likelihood step, which relies on an estimate of the probability of being selected in population 1 as a function of the covariates. The resulting targeted maximum likelihood estimator is double robust and efficient. That is, with our formulation we make it possible to use state of the art statistical methodology in causal inference to obtain fully efficient and double robust estimators of the causal effect of an intervention assigned at the population level.

We show in the accompanying technical report 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 can potentially be used to 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.

We extend our results to identification and estimation of the effects of community-based interventions using individually matched two community cohort designs, and investigate practical and theoretical efficiency gains under this design compared to one in which individuals are sampled independently. We use general results on efficient influence curves and targeted maximum likelihood estimation for case-control sampling as established in van der Laan [33] 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 treated population, showing the strong potential benefits of matching for the purpose of this target causal parameter.

Finally, we generalize our identifiability theorems, efficient influence curve, and estimators to the case that one assigns two possible interventions to K communities. 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 $\sum_k n_k$, the pooled sample size across all communities.

If a large number of communities is sampled from some target population of communities, so that asymptotics in the number of communities is sensible, then one can obtain identifiability of the causal effects of community level interventions under the usual

randomization and positivity assumption known from the classical causal inference literature. In such a case, one need not rely on the strong identifiability assumptions assumed in this article.

The targeted maximum likelihood estimators presented here for a fixed number of K communities are still consistent and asymptotically linear if the intervention is randomly assigned at the community level, as in cluster randomized trial designs. They do not involve adjustment by environmental factors, however, and are, therefore, somewhat inefficient. On the other hand, even in the case that one aims to identify a causal effect for a target population of communities, the targeted maximum likelihood estimator presented here remains of interest when the number of communities is small, the number of sampled individuals is large, and the exclusion restriction assumption holds up to a reasonable approximation. We refer to the technical report van der Laan [34] for results for this TMLE with respect to a causal target parameter defined with respect to a target population of communities in the context that K converges to infinity, in the case where the exclusion restriction assumption does not necessarily hold.

Acknowledgments

Mark van der Laan and Wenjing Zheng are supported by NIH grant R01 AI074345-06. Maya Petersen is a recipient of a Doris Duke Clinical Scientist Development Award.

References

1. Oakes JM. The (mis)estimation of neighborhood effects: causal inference for a practicable social epidemiology. *Soc Sci Med*. 2004; 58(10):1953–1960. [PubMed: 15020010]
2. Small DS, Ten Have TR, Rosenbaum PR. Randomization inference in a group-randomized trial of treatments for depression. *J Am Stat Assoc*. 2008; 1030:271–79.
3. Hanse B, Bowers J. Attributing effects to a cluster-randomized get-out-the-vote campaign. *J Am Stat Assoc*. 2009; 1040:873–85.
4. Imai K, King G, Nall C. The essential role of pair matching in cluster-randomized experiments, with application to the mexican universal health insurance evaluation. *Stat Sci*. 2009; 240:29–53.
5. Varnell SP, Murray DM, Janega JB, Blitstein JL. Design and analysis of group-randomized trials: a review of recent practices. *Am J Public Health*. 2004; 94:393–99. [PubMed: 14998802]
6. Atienza A, King A. Community-based health intervention trials: an overview of methodological issues. *Epidemiol Rev*. 2002; 24:72–9. [PubMed: 12119859]
7. Kenward MG, Roger JH. *Biometrics*. 1997; 53:983–97. [PubMed: 9333350]
8. Pan W, Wall MM. Small-sample adjustments in using the sandwich variance estimator in generalized estimating equations. *Stat Med*. 2002; 21:1429–41. [PubMed: 12185894]
9. Skene S, Kenward MG. The analysis of very small samples of repeated measurements I: an adjusted sandwich estimator. *Stat Med*. 2010a; 29:2825–37. [PubMed: 20839367]
10. Skene S, Kenward MG. The analysis of very small samples of repeated measurements I: a modified box correction. *Stat Med*. 2010b; 29:2837–56.
11. Zhang K, Small DS. Comment: the essential role of pair matching in cluster-randomized experiments, with application to the mexican universal health insurance evaluation. *Stat Sci*. 2009; 24:59–64.
12. Cox, DR. *The planning of experiments*. New York: Wiley; 1958.
13. Rubin DB. Bayesian inference for causal effects: the role of randomization. *Ann Stat*. 1978; 6:34–58.

14. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika*. 1983; 70:41–55.
15. Holland PW. Statistics and causal inference. *J Am Stat Assoc*. 1986; 81:945–60.
16. Hong B, Raudenbush S. Evaluating kindergarten retention policy. *J Am Stat Assoc*. 2006; 101:901–10.
17. Sobel M. What do randomized studies of housing mobility demonstrate? *J Am Stat Assoc*. 2006; 101:1398–407.
18. Halloran ME, Struchiner CJ. Causal inference in infectious diseases. *Epidemiology*. 1995; 60:142–51. [PubMed: 7742400]
19. Hudgens MG, Halloran ME. Towards causal inference with interference. *J Am Stat Assoc*. 2008; 103:832–42. [PubMed: 19081744]
20. VanderWeele TJ, Tchetgen Tchetgen EJ. Bounding the infectiousness effect in vaccine trials. *Epidemiology*. 2011a; 22:686–93. [PubMed: 21753730]
21. VanderWeele TJ, Tchetgen Tchetgen EJ. Effect partitioning under interference for two-stage randomized vaccine trials. *Stat Probab Lett*. 2011b; 81:861–69. [PubMed: 21532912]
22. Rosenbaum P. Interference between units in randomized experiments. *J Am Stat Assoc*. 2007; 102:191–200.
23. Tchetgen Tchetgen EJ, VanderWeele TJ. On causal inference in the presence of interference. *Statistical Methods in Medical Research – Special Issue on Causal Inference*. 2012; 21:55–75.
24. Sinclair B, McConnell M, Green D. Detecting spillover in social networks: design and analysis of multi-level experiments. *Am J Political Sci*. 2012; 56(4):1055–1069.
25. Shalizi CR, Thomas AC. Homophily and contagion are generically confounded in observational social network studies. *SociolMethods Res*. 2011; 40:211–39.
26. Manski CF. Identification of endogenous social effects: the reflection problem. *Rev Econ Stud*. 1993; 60:531–42.
27. Graham B. Identifying social interactions through conditional variance restrictions. *Econometrica*. 2008; 76:643–60.
28. Graham B, Imbens GW, Ridder G. Measuring the effects of segregation in the presence of social spillovers: a nonparametric approach. Technical report. 2010
29. Pearl, J. *Causality: models, reasoning and inference*. 2. New York: Cambridge University Press; 2009.
30. Varnell SP, Murray DM, Baker W. An evaluation of analysis options for the one-group-per-condition design: can any of the alternatives overcome the problems inherent in this design? *Eval Rev*. 2001; 25:440–53. [PubMed: 11480307]
31. Polley, EC.; van der Laan, MJ. Technical Report 266. Division of Biostatistics, University of California; Berkeley: 2010. Super learner in prediction.
32. Rose, S.; van der Laan, MJ. Simple optimal weighting of cases and controls in case-control studies. *Int J Biostat*. 2008. Available at: <http://www.bepress.com/ijb/vol4/iss1/19/>
33. van der Laan, MJ. Estimation based on case-control designs with known prevalence probability. *Int J Biostat*. 2008. <http://www.bepress.com/ijb/vol4/iss1/17/>
34. van der Laan, MJ. Working paper. Vol. 268. U.C. Berkeley Division of Biostatistics Working Paper Series; Jun. 2010 Estimation of causal effects of community based interventions. <http://www.bepress.com/ucbbiostat/paper268>
35. van der Laan MJ, Rubin DB. Targeted maximum likelihood learning. *Int J Biostat*. 2006; 2 Available at <http://www.degruyter.com/view/j/ijb.2006.2.1/ijb.2006.2.1.1043/ijb.2006.2.1.1043.xml?format=INT>.
36. van der Laan, MJ.; Rose, S. Targeted learning. New York: Springer; 2011.
37. van der Laan MJ, Gruber S. Collaborative double robust penalized targeted maximum likelihood estimation. *Int J Biostat*. 2010; 6 Available at <http://www.degruyter.com/view/j/ijb.2010.6.1/ijb.2010.6.1.1181/ijb.2010.6.1.1181.xml?format=INT>.
38. Gruber S, van der Laan MJ. An application of collaborative targeted maximum likelihood estimation in causal inference and genomics. *Int J Biostat*. 2010:60.
39. van der Laan MJ, Polley EC, Hubbard AE. Super learner. *Stat Appl Genet Mol Biol*. 2007; 60

40. Rose, S.; van der Laan, MJ. Why match? Investigating matched case-control study designs with causal effect estimation. *Int J Biostat*. 2009. Available at: <http://www.bepress.com/ijb/vol5/iss1/1/>
41. Gruber, S.; van der Laan, MJ. Technical report 265. Division of Biostatistics, University of California; Berkeley: May. 2010 A targeted maximum likelihood estimator of a causal effect on a bounded continuous outcome.

Table 1

Double Robustness: 1-to-1 Matched cohort design. Efficiency bounds are 1.139 and 0.1139 for $n_j = 100$ and 1,000, respectively.

$n_1 =$	100				1,000			
	mean	$\ bias\ $	var	MSE	mean	$\ bias\ $	var	MSE
$\Psi(Q_{unadj})$	15.79	7.68	44.15	103.08	15.69	7.78	4.71	65.18
$\Psi(\hat{Q}_{unadj}^*(g_c))$	22.93	0.54	3.62	3.91	23.42	0.05	0.25	0.25
$\Psi(Q_m)$	28.59	5.12	5.64	31.89	28.62	5.15	0.56	27.04
$\Psi(\hat{Q}_m^*(g_c))$	23.53	0.06	3.19	3.19	23.50	0.03	0.32	0.32
$\Psi(Q_c)$	23.45	0.02	1.14	1.14	23.48	0.01	0.12	0.12
$\Psi(\hat{Q}_c^*(g_m))$	23.44	0.03	1.15	1.15	23.48	0.01	0.12	0.12
$\Psi(\hat{Q}_c^*(g_c))$	23.45	0.02	1.13	1.13	23.48	0.01	0.12	0.12

Table 2

Double Robustness: Independent design. Efficiency bounds are 1.337 and 0.134 for $n_j = 100$ and 1,000, respectively.

$n_j =$	100				1,000			
	mean	$\ bias\ $	var	MSE	mean	$\ bias\ $	var	MSE
$\Psi(\hat{Q}_{unadj})$	36.61	13.14	45.57	218.23	36.66	13.19	4.88	178.88
$\Psi(\hat{Q}_{unadj}^*(g_c))$	22.96	0.51	44.82	45.08	23.51	0.04	3.69	3.69
$\Psi(\hat{Q}_m)$	28.68	5.21	5.16	32.27	28.58	5.11	0.47	26.58
$\Psi(\hat{Q}_m^*(g_c))$	23.50	0.04	7.31	7.31	23.46	0.01	0.60	0.60
$\Psi(\hat{Q}_c)$	23.46	0.01	1.31	1.31	23.48	0.01	0.14	0.14
$\Psi(\hat{Q}_c^*(g_m))$	23.45	0.02	1.34	1.34	23.48	0.01	0.14	0.14
$\Psi(\hat{Q}_c^*(g_c))$	23.45	0.02	1.34	1.34	23.48	0.01	0.14	0.14

Table 3

Ratio of the estimated variances of non-targeted and targeted substitution estimators, under independent versus matched sampling ($\widehat{var}(\psi_{match})/ \widehat{var}(\psi_{indep})$), for $n_1 = n_2 = 5000$, and for each of the nine data generating distributions. The first two columns indicate the extent to which the matching variable predicts the intervention A and the outcome, Y , respectively.

A	Y	$\Psi(Q_{indep})$	$\Psi(Q_{indep}^*(g_c))$	$\Psi(Q_m)$	$\Psi(Q_m^*(g_c))$	$\Psi(Q_c)$	$\Psi(Q_c^*(g_c))$
Low	Low	1.04	0.90	1.11	1.0	1.04	1.04
Low	Medium	1.03	0.96	0.90	1.1	1.03	1.03
Low	High	1.10	1.03	0.93	1.0	1.02	1.02
Medium	Low	1.00	2.11	1.05	1.1	1.05	1.06
Medium	Medium	0.92	2.07	0.98	1.1	0.96	0.98
Medium	High	1.00	1.94	0.87	1.0	0.95	0.96
High	Low	1.05	11.75	0.89	1.9	1.22	1.24
High	Medium	0.99	11.57	0.88	1.9	1.19	1.22
High	High	1.00	15.30	0.82	1.9	1.15	1.15

Table 4

$\frac{\hat{V}_{ar}(D^*(P_0))}{2\hat{V}_{ar}(D^{m,*}(P_0))}$, relative asymptotic efficiency of the independent versus matched designs, at $g = g_0$, and $Q = Q_0$, for each of the nine data generating distributions.

Extent to which W_2 predicts Y	Extent to which W_2 predicts A		
	Low	Medium	High
Low	0.997	1.024	1.231
Medium	0.998	1.025	1.220
High	1.001	1.016	1.174

Table 5

$\frac{\hat{V}ar(D^*(Q_m, g_0, \psi_0))}{2\hat{V}ar(D^{m,*}(Q_m, g_0, \psi_0))}$, relative asymptotic efficiency of the independent versus matched designs, at $g = g_0$, and $Q = Q_m$, for each of the nine data generating distributions.

Extent to which W_2 predicts Y	Extent to which W_2 predicts A		
	Low	Medium	High
Low	1.000	1.125	2.139
Medium	1.009	1.138	2.262
High	1.000	1.124	2.543

Table 6

$\frac{\hat{V}ar(D^*(Q_{unadj}, g_0, \psi_0))}{2\hat{V}ar(D^{m,*}(Q_{unadj}, g_0, \psi_0))}$ relative asymptotic efficiency of the independent versus matched designs, at $g = g_0$, and $Q = Q_{unadj}$, for each of the nine data generating distributions.

Extent to which W_2 predicts Y	Extent to which W_2 predicts A		
	Low	Medium	High
Low	1.028	1.118	1.895
Medium	1.026	1.096	1.814
High	1.026	1.059	1.986