Biometrika (2022), **109**, 4, *pp.* 1165–1172 https://doi.org/10.1093/biomet/asac008
Advance Access publication 8 February 2022

Average direct and indirect causal effects under interference

BY YUCHEN HU

Department of Management Science and Engineering, Stanford University, 475 Via Ortega, Stanford, California 94305, U.S.A.
yuchenhu@stanford.edu

SHUANGNING LI

Department of Statistics, Stanford University, 390 Jane Stanford Way, Stanford, California 94305, U.S.A.
lsn@stanford.edu

AND STEFAN WAGER

Stanford Graduate School of Business, 655 Knight Way, Stanford, California 94305, U.S.A.
swager@stanford.edu

SUMMARY

We propose a definition for the average indirect effect of a binary treatment in the potential outcomes model for causal inference under cross-unit interference. Our definition is analogous to the standard definition of the average direct effect and can be expressed without needing to compare outcomes across multiple randomized experiments. We show that the proposed indirect effect satisfies a decomposition theorem stating that in a Bernoulli trial, the sum of the average direct and indirect effects always corresponds to the effect of a policy intervention that infinitesimally increases treatment probabilities. We also consider a number of parametric models for interference and find that our nonparametric indirect effect remains a natural estimand when re-expressed in the context of these models.

Some key words: Causal inference; Interference; Potential outcome; Randomized trial.

1. INTRODUCTION

The classical way of analysing randomized trials, following Neyman (1923) and Rubin (1974), centres on the average treatment effect defined using potential outcomes. Given a sample of $i = 1, \ldots, n$ units used to study the effect of a binary treatment $W_i \in \{0, 1\}$, we posit potential outcomes $Y_i(0), Y_i(1) \in \mathbb{R}$ corresponding to the outcomes we would have measured had we assigned the ith unit to control and to treatment, respectively, i.e., we observe $Y_i = Y_i(W_i)$. We then proceed by arguing that the sample average treatment effect

$$\tau_{\text{ATE}} = \frac{1}{n} \sum_{i=1}^{n} \{Y_i(1) - Y_i(0)\} \tag{1}$$

admits a simple unbiased estimator under random assignment of treatment.

One limitation of this classical approach is that it rules out interference and instead introduces an assumption that the observed outcome for any given unit does not depend on the treatments assigned to other units, i.e., Y_i is not affected by W_j for any $j \neq i$ (Halloran & Struchiner, 1995). However, in a wide variety of applied settings, such interference effects not only exist, but are often of considerable scientific interest (Sacerdote, 2001; Miguel & Kremer, 2004; Bakshy et al., 2012; Bond et al., 2012; Cai et al., 2015;

Rogers & Feller, 2018). For example, in an education setting, it may be of interest to understand how a pedagogical innovation affects not just certain targeted students, but also their peers. This has led to a recent surge of interest in methods for studying randomized trials under interference (Hudgens & Halloran, 2008; Tchetgen Tchetgen & VanderWeele, 2012; Manski, 2013; Aronow & Samii, 2017; Eckles et al., 2017; Leung, 2020; Li & Wager, 2020; Sävje et al., 2021).

A major difficulty in working under interference is that one no longer has a single obvious average effect parameter to target as in (1). In the general setting, each unit now has 2^n potential outcomes corresponding to every possible treatment combination assigned to the n units, and these can be used to formulate effectively innumerable possible treatment effects that can arise from different assignment patterns. As discussed further in § 3, the existing literature has mostly side-stepped this issue by framing the estimand in terms of specific policy interventions. However, this paradigm does not provide researchers with simple, nonparametric and agnostic average causal estimands that can be studied without spelling out a specific policy intervention of interest.

In this paper we study a pair of averaging causal estimands, the average direct and indirect effects, that are valid under interference and yet, unlike existing targets, can be defined and estimated using a single experiment and do not need to be defined in terms of hypothetical policy interventions. Qualitatively, the average direct effect measures the extent to which, in a given experiment and on average, the outcome Y_i of a unit is affected by its own treatment W_i; meanwhile, the average indirect effect measures the responsiveness of Y_i to treatments W_j given to other units $j \neq i$.

The average direct effect we consider is standard and has recently been discussed by a number of authors, including VanderWeele & Tchetgen Tchetgen (2011) and Sävje et al. (2021). Our definition of the average indirect effect is, to the best of our knowledge, new and is the main contribution of this paper. We follow this definition with a number of results to validate it. In particular, we prove a universal decomposition theorem which says that in a Bernoulli trial the sum of the average direct and indirect effects can always be interpreted as the total effect of an intuitive policy intervention. We also interpret these estimands in the context of a number of parametric models for interference considered by practitioners.

2. Treatment effects under interference

We study different experimental designs using the potential outcomes model. The main difference between a setting with interference and the standard Neyman–Rubin model is that potential outcomes for the ith unit may also depend on the intervention given to the jth unit with $j \neq i$ (e.g., Hudgens & Halloran, 2008; Aronow & Samii, 2017). For convenience, we use the shorthand $Y_i(w_j = x; W_{-j})$ to denote the potential outcome we would observe for the ith unit if we were to assign the jth unit to treatment status $x \in \{0, 1\}$ and maintain all units, but the jth at their realized treatments $W_{-j} \in \{0, 1\}^{n-1}$. Expectations E are over the treatment assignment only; potential outcomes are held fixed.

Assumption 1. For units $i = 1, \ldots, n$ there are potential outcomes $Y_i(w) \in \mathbb{R}$, $w \in \{0, 1\}^n$, such that given a treatment vector $W \in \{0, 1\}^n$ we observe outcomes $Y_i = Y_i(W)$.

Definition 1. *Under Assumption 1, the average direct effect of a binary treatment is*

$$\tau_{\text{ADE}} = \frac{1}{n} \sum_{i=1}^{n} E\left[Y_i(w_i = 1; W_{-i}) - Y_i(w_i = 0; W_{-i})\right].$$

Definition 2. *Under Assumption 1, the average indirect effect of a binary treatment is*

$$\tau_{\text{AIE}} = \frac{1}{n} \sum_{i=1}^{n} \sum_{j \neq i} E\left[Y_j(w_i = 1; W_{-i}) - Y_j(w_i = 0; W_{-i})\right].$$

The definition of the direct effect τ_{ADE} is standard. It follows from averaging $Y_i(w_i = 1; W_{-i}) - Y_i(w_i = 0; W_{-i})$, referred to as the direct causal effect by Halloran & Struchiner (1995). Recently Sävje et al.

(2021) provided an in-depth discussion of this estimand. This estimand measures the average effect of an intervention W_i on the unit being intervened on, while marginalizing over the rest of the treatment assignments. In a study without interference, τ_{ADE} matches the usual average treatment effect (1).

Meanwhile, our definition of the indirect effect is an immediate formal generalization of τ_{ADE} to cross-unit treatment effects. It measures the average effect of an intervention W_i on all units except the one being intervened on, again marginalizing over the rest of the process. More precisely, the term $E\{Y_j(w_i = 1; W_{-i}) - Y_j(w_i = 0; W_{-i})\}$ is the effect of changing unit i's treatment on the outcome of unit j. Thus the sum $\sum_{j \neq i} E\{Y_j(w_i = 1; W_{-i}) - Y_j(w_i = 0; W_{-i})\}$ would correspond to the aggregate effect of unit i's treatment on all the other units. Then the defined average indirect effect τ_{AIE} corresponds to the average of the effects of units' treatments on other units.

The definition of τ_{AIE} formally mirrors that of τ_{ADE}, and in the no-interference case we clearly have $\tau_{\text{AIE}} = 0$. As a first step towards validating the definition of τ_{AIE} under nontrivial interference, we consider the average overall effect induced by adding τ_{ADE} and τ_{AIE}, which aggregates the marginalized effect of all treatments on all outcomes. We then prove that in a Bernoulli design, this matches the policy effect of infinitesimally increasing each unit's treatment probability. We use the term Bernoulli design to refer to an experiment where there is a deterministic vector $\pi \in (0, 1)^n$ such that the treatments W_i are generated as $W_i \sim \text{Ber}(\pi_i)$ for all $i = 1, \ldots, n$, independently of each other and of the potential outcomes $\{Y_i(w)\}$. For a Bernoulli design with treatment probabilities $\pi \in [0, 1]^n$, we write $E_\pi(\cdot)$ for expectations over the random treatment assignment, and we write $\tau_{\text{ADE}}(\pi)$, $\tau_{\text{AIE}}(\pi)$ and $\tau_{\text{AOE}}(\pi)$ for the corresponding direct, indirect and overall effects.

DEFINITION 3. *Under Assumption 1, the average overall effect of a binary treatment is*

$$\tau_{\text{AOE}} = \tau_{\text{ADE}} + \tau_{\text{AIE}} = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{n} E\big\{Y_j(w_i = 1; W_{-i}) - Y_j(w_i = 0; W_{-i})\big\}.$$

DEFINITION 4. *Under Assumption 1 and in a Bernoulli design, the infinitesimal policy effect is*

$$\tau_{\text{INF}}(\pi) = 1 \cdot \nabla_\pi E_\pi \left(\frac{1}{n} \sum_{i=1}^{n} Y_i \right) = \sum_{k=1}^{n} \frac{\partial}{\partial \pi_k} E_\pi \left(\frac{1}{n} \sum_{i=1}^{n} Y_i \right).$$

THEOREM 1. *Under Assumption 1 and in a Bernoulli design,* $\tau_{\text{AOE}}(\pi) = \tau_{\text{INF}}(\pi)$.

By connecting our abstract notions of direct, indirect and overall effects to the effect of a concrete policy intervention, Theorem 1 provides an alternative lens on our definition of the indirect effect. Suppose, for example, that a researcher knew they wanted to study nudge interventions, the total effect of which is $\tau_{\text{INF}}(\pi)$, and was also committed to the standard definition of the average direct effect given in Definition 1. Then it would be natural to define an indirect effect as $\tau_{\text{INF}}(\pi) - \tau_{\text{ADE}}(\pi)$, i.e., to characterize as indirect any effect of the nudge intervention that is not captured by the direct effect; this is, for example, the approach implicitly taken in Heckman et al. (1998). From this perspective, Theorem 1 could be seen as showing that these two possible definitions of the indirect effect in fact match, i.e., that $\tau_{\text{AIE}}(\pi) = \tau_{\text{INF}}(\pi) - \tau_{\text{ADE}}(\pi)$. We emphasize that Theorem 1 is a direct consequence of Bernoulli randomization and holds conditionally on any realization of the potential outcomes $\{Y_i(w)\}$.

We refer to $\tau_{\text{INF}}(\pi)$ as a policy effect because in the ideal situation where one has access to observed outcomes Y_i for different randomization probabilities π, $\tau_{\text{INF}}(\pi)$ is a quantity that could be measured by averaging observed outcomes Y_i for different π. If treatment assignment probabilities are constant, i.e., there is a $\pi_0 \in (0, 1)$ such that $\pi_i = \pi_0$ for all $i = 1, \ldots, n$, then $\tau_{\text{INF}}(\pi)$ takes a particularly simple form, $\tau_{\text{INF}}(\pi) = (\text{d}/\text{d}\pi_0) E_{\pi_0}\big(n^{-1} \sum_{i=1}^{n} Y_i\big)$. Infinitesimal policy effects as defined above are prevalent in the social sciences owing to their ease of interpretation and desirable analytical properties; see, for example, Chetty (2009), Carneiro et al. (2010) and references therein. Wager & Xu (2021) discussed welfare implications for a social planner who uses analogous infinitesimal policy effects to optimize a system via gradient-based methods.

3. ALTERNATIVE DEFINITIONS AND RELATED WORK

Various other average causal effect estimands have recently been discussed in the literature. In the case of direct effects, the main alternative to Definition 1 is the following proposal of Hudgens & Halloran (2008) that relies on conditional expectations:

$$\tau_{\text{HH, DE}} = \frac{1}{n}\sum_{i=1}^{n}\{E(Y_i \mid W_i = 1) - E(Y_i \mid W_i = 0)\}.$$

In a Bernoulli design, $\tau_{\text{HH, DE}} = \tau_{\text{ADE}}$. However, in other designs, such as completely randomized designs or stratified designs, these two estimands do not match. As discussed in VanderWeele & Tchetgen Tchetgen (2011) and Sävje et al. (2021), a major drawback of $\tau_{\text{HH, DE}}$ is that it conflates the effect of setting $w_i = x$ on the ith unit's outcome and the effect of setting $w_i = x$ on the distribution of W_{-i}. In particular, in completely randomized experiments, it is possible to have $\tau_{\text{HH, DE}} \neq 0$ even when $Y_i(w_i = 1, w_{-i}) = Y_i(w_i = 0, w_{-i})$ for all units and all possible treatment assignments. In contrast, the τ_{ADE} in Definition 1 has a robust causal interpretation as a direct effect.

Meanwhile, as discussed in the introduction, most available notions of indirect effects rely on explicit comparisons between two overall treatment assignment strategies. For example, Hudgens & Halloran (2008) and VanderWeele & Tchetgen Tchetgen (2011) proposed a number of indirect effect estimands that, in the case of comparing two Bernoulli trials with randomization probabilities π and π', reduce to

$$\tau_{\text{IE}}(\pi, \pi') = \frac{1}{n}\sum_{i=1}^{n}\big[E_\pi\{Y_i(w_i = 0; W_{-i})\} - E_{\pi}\{Y_i(w_i = 0; W_{-i})\}\big].$$

In the case of non-Bernoulli trials, there are subtleties analogous to the ones noted above; see VanderWeele & Tchetgen Tchetgen (2011) for an in-depth discussion. Although $\tau_{\text{IE}}(\pi, \pi')$ is an interesting quantity to consider if we can run many independent experiments that test different overall treatment levels, unlike τ_{AIE} it does not enable a researcher to describe indirect effects in a single randomized study.

Another popular way of capturing indirect effects is via the exposure mapping approach developed by Aronow & Samii (2017). The main idea is to assume existence of functions $h_i : \{0, 1\}^n \to \{1, \ldots, K\}$ such that potential outcomes $Y_i(w)$ depend on w only via the compressed representation $h_i(w)$, i.e., $Y_i(w) = Y_i(w')$ whenever $h_i(w) = h_i(w')$; see also Karwa & Airoldi (2018), Leung (2020) and Sävje (2021) for further discussions and extensions. One can then consider estimators of averages of potential outcome types and define treatment effects in terms of their contrasts,

$$\mu(k) = \frac{1}{n}\sum_{i=1}^{n} E\{Y_i \mid h_i(W_i) = k\}, \quad \tau(k, k') = \mu(k') - \mu(k) \quad (1 \leqslant k \neq k' \leqslant K).$$

Definitions of this type are again conceptually attractive and sometimes enable us to very clearly express the answer to a natural policy question, see, e.g., Basse et al. (2019). However, they again require the analyst to consider specific policy interventions to be able to even talk about indirect effects, and they can also be unwieldy to use as the number K of possible exposure types gets large.

Closest to the definition of τ_{AIE} is the average marginalized response of Aronow et al. (2021). They considered a setting where treatments are assigned to points in a geographic space and sought to estimate the average effect of treatment at an intervention point on outcomes at points that are a distance d away,

$$\tau_{\text{AMR}}(d; \pi) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j \in \mathcal{S}_i(d)} \frac{E_\pi\{Y_j(w_i = 1; W_{-i}) - Y_j(w_i = 0; W_{-i})\}}{|\mathcal{S}_i(d)|}, \quad \mathcal{S}_i(d) = \{j : \Delta(i, j) = d\},$$

where $\Delta(i, j)$ measures the distance between points i and j. This circle average bears a resemblance to Definition 2 in the sense that both are marginalized over variation in W_{-i} while holding the treatment W_i

fixed. However, one key difference is the normalization factor $|\mathcal{S}_i(d)|^{-1}$ used in τ_{AMR}. Adding a similar normalization to τ_{AIE} would invalidate Theorem 1.

Remark 1. Our definition of $\tau_{\text{AIE}}(\pi)$ is normalized by n, not by the total number of summands $n(n-1)$, and one can ask whether this scaling is always the most natural one. We argue below that in several popular models $\tau_{\text{AIE}}(\pi)$ coincides with interesting and interpretable quantities, and converges as the number n of units goes to infinity. In other models, however, our $1/n$ scaling may not be the best choice. For example, if we have a data-generating distribution with $Y_i(w) = (\sum_{j=1}^n w_j - n\pi)/\{n\pi(1-\pi)\}^{1/2}$, then the observed outcomes Y_i will have a standard normal marginal distribution, but $\tau_{\text{AIE}} = \sqrt{n}$ will diverge. It should also be noted, however, that in this example $n^{-1}\sum_{i=1}^n Y_i$ does not concentrate.

4. Models for interference

Our discussion so far has focused on an abstract specification where direct and indirect effects are defined via various marginalized contrasts between potential outcomes. Much of the existing applied work on causal inference under interference, however, has focused on simpler parametric specifications that, for instance, connect outcomes to treatments via a linear model. The purpose of this section is to examine our abstract, nonparametric definition of the indirect effect in Definition 2, and to confirm that it still corresponds to an estimand we would want to interpret as an indirect effect when we restrict our attention to simpler parametric models. Below, we do so in three examples. An extended study of τ_{ADE} and τ_{AIE} in a marketplace model where interference arises via equilibrium price formation can be found in Munro et al. (2021). The claimed expressions for τ_{ADE} and τ_{AIE} are derived in the Supplementary Material.

Example 1. In studying the spillover effects of training sessions on insurance purchase, Cai et al. (2015) used a network model: there is an edge matrix $E_{ij} \in \{0,1\}$ such that W_j can affect Y_i only if the corresponding units are connected by an edge, i.e., $E_{ij} = 1$. They then considered a linear-in-means model parameterized in terms of this network. For our purpose, we focus on a simple variant of the model of Cai et al. (2015) considered in Leung (2020), where only the effects of ego's treatment and the proportion of treated neighbours are considered as covariates. This results in a linear model induced by the structural equation

$$Y_i = \beta_1 + \beta_2 W_i + \beta_3 \frac{\sum_{j\neq i} E_{ij} W_j}{\sum_{j\neq i} E_{ij}} + \varepsilon_i, \quad E(\varepsilon_i \mid W) = 0. \tag{2}$$

In other words, the probability of insurance purchase is modelled as a linear function of whether the farmer attends the insurance training sessions and the proportion of friends who attend the session. The relation (2) should be taken as a structural model, meaning that we can generate potential outcomes $Y_i(w)$ by plugging candidate assignment vectors w into (2), i.e., $Y_i(w) = \beta_1 + \beta_2 w_i + \beta_3 \sum_{j\neq i} E_{ij} w_j / \sum_{j\neq i} E_{ij} + \varepsilon_i$ for all $w \in \{0,1\}^n$. With this model, it can be shown that under Assumption 1 we have $\tau_{\text{ADE}} = \beta_2$ and $\tau_{\text{AIE}} = \beta_3$, i.e., the estimands from Definitions 1 and 2 map exactly to the parameters in model (2) regardless of the experimental design.

Example 2. The model in Example 1 assumes that the ith unit responds in the same way to treatment assigned to any of its neighbours. However, this restriction may be implausible in many situations; for example, in social networks there is evidence that some ties are stronger than others and that peer effects are greater along strong ties (Bakshy et al., 2012). A natural generalization of Example 1 that allows for variable-strength ties uses a saturated structural linear model

$$Y_i = \alpha_i + \beta_i W_i + \sum_{j\neq i} \nu_{ij} W_j + \varepsilon_i, \quad E(\varepsilon_i \mid W) = 0, \tag{3}$$

which allows for unit-specific direct and indirect effects. Here the individual parameters in this model are not identifiable; however, under (3), $\tau_{\text{ADE}} = n^{-1}\sum_{i=1}^n \beta_i$ and $\tau_{\text{AIE}} = n^{-1}\sum_{i=1}^n \sum_{j\neq i} \nu_{ij}$, i.e., our estimands can be understood as averages of the unit-level parameters, again regardless of the design.

Example 3. In studying the effects of persuasion campaigns or other types of messaging, one might assume that people respond most strongly if they receive a communication directly addressed to them, but may also respond if a member of their neighbourhood or household gets a communication. This assumption can be formalized in terms of a model where each unit has four potential outcomes:

$$Y_i = \begin{cases} Y_i(\text{treated \& exposed}), & W_i = 1 \text{ and } i \text{ has a treated neighbour}, \\ Y_i(\text{treated}), & W_i = 1 \text{ but } i \text{ has no treated neighbours}, \\ Y_i(\text{exposed}), & W_i = 0 \text{ but } i \text{ has a treated neighbour}, \\ Y_i(\text{none}), & W_i = 0 \text{ and } i \text{ has no treated neighbours}. \end{cases}$$

Models of this type have been considered by Sinclair et al. (2012) in studying voter mobilization and by Basse & Feller (2018) and Basse et al. (2019) for studying anti-absenteeism interventions. Natural treatment effect parameters to consider following (3) include the average self-treatment and spillover effects

$$\tau_{\text{SELF},1} = \frac{1}{n}\sum_{i=1}^{n}\{Y_i(\text{treated \& exposed}) - Y_i(\text{exposed})\}, \quad \tau_{\text{SELF},0} = \frac{1}{n}\sum_{i=1}^{n}\{Y_i(\text{treated}) - Y_i(\text{none})\},$$

$$\tau_{\text{SPILL},1} = \frac{1}{n}\sum_{i=1}^{n}\{Y_i(\text{treated \& exposed}) - Y_i(\text{treated})\}, \quad \tau_{\text{SPILL},0} = \frac{1}{n}\sum_{i=1}^{n}\{Y_i(\text{exposed}) - Y_i(\text{none})\}.$$

Unlike in the previous two examples, the connection of τ_{ADE} and τ_{AIE} with τ_{SELF} and τ_{SPILL} differs substantially across experimental designs, especially when the design introduces correlation between the units. For illustration, we study this example with a multi-stage completely randomized design considered in previous works (Sinclair et al., 2012; Basse & Feller, 2018; Basse et al., 2019). In particular, we focus on the case where there are a total of n/m clusters of size m. In the first stage, $\rho \times n/m$ clusters are assigned to treatment and $(1-\rho) \times n/m$ clusters are assigned to control; in the second stage, a single unit in each treated cluster is randomly chosen to be treated, and all the other units are assigned to control. We can then calculate the marginal distribution of W_{-i} and obtain

$$\tau_{\text{ADE}} = \left(\rho - \frac{\rho}{m}\right)\tau_{\text{SELF},1} + \left(1 - \rho + \frac{\rho}{m}\right)\tau_{\text{SELF},0},$$
$$\tau_{\text{AIE}} = (m-1)\left\{\frac{\rho}{m}\tau_{\text{SPILL},1} + \left(1 - \rho + \frac{\rho}{m}\right)\tau_{\text{SPILL},0}\right\}.$$

Therefore, our estimands can be regarded as weighted averages of the treatment effect parameters. Moreover, when the cluster size m is large, τ_{ADE} is approximately the average of $\tau_{\text{SELF},1}$ and $\tau_{\text{SELF},0}$ weighted by the assignment probability during the first stage, while τ_{AIE} is approximately $\tau_{\text{SPILL},0}$ times the probability of being assigned to the control group during the first stage, and the factor $m-1$ simply accounts for the fact that any treatment will spread spillover effects to $m-1$ neighbours.

5. DISCUSSION

Our definition of τ_{AIE} also has the potential to help synthesize nonparametric and model-based approaches to interference by providing a shared estimand that can be studied from both perspectives: as discussed in § 4, although τ_{AIE} is defined in terms of a generic potential outcomes model, it is also a natural estimand in a number of different structural models. Munro et al. (2021) have pursued this agenda further, showing that in a marketplace governed by a general equilibrium model where prices mediate interference, τ_{AIE} can be expressed in terms of familiar economic quantities such as price elasticities.

One challenge is that our estimands will in general depend on the design. Figure 1 illustrates this phenomenon in a Bernoulli experiment by plotting $\tau_{\text{ADE}}(\pi)$ and $\tau_{\text{AIE}}(\pi)$ as functions of π in the following

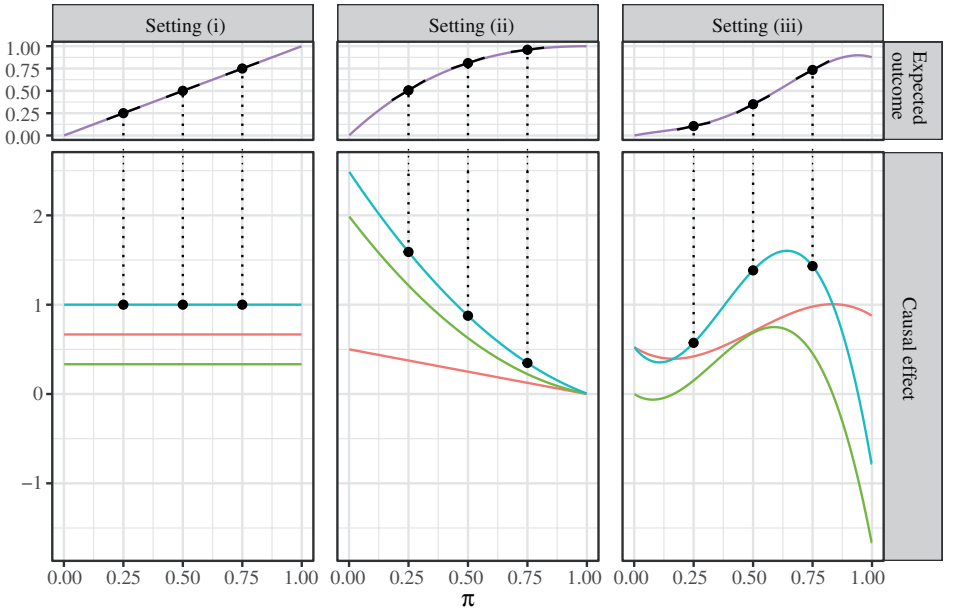

Fig. 1. Plots of τ_{ADE} (red), τ_{AIE} (green), τ_{INF} (blue) and the expected potential outcome $E_\pi(Y_i)$ (purple) as functions of the treatment probability π. The slopes of tangent line segments on the purple curve, which represent the derivative of $E_\pi(Y_i)$ at those points, are the same as the values on the blue curve, τ_{INF}. Theorem 1 establishes that $\tau_{\text{INF}} = \tau_{\text{ADE}} + \tau_{\text{AIE}}$. In the plots, the blue curve corresponds to the sum of the red curve and the green curve. We consider the three settings in (4), where in all cases we assume constant treatment assignment probability $\pi_i = \pi_0$ and take the number of neighbours to be $\sum_{i \neq j} E_{ij} = 100$.

three structural models with constant treatment probability π:

$$\begin{aligned} &\text{(i)}\ \ Y_i = \frac{\sum_{i\neq j} E_{ij} W_j}{300} + \frac{2W_i}{3} + \varepsilon_i, \quad \text{(ii)}\ \ Y_i = 1 - \left(1 - \frac{\sum_{i\neq j} E_{ij} W_j}{\sum_{i\neq j} E_{ij}}\right)^2 \left(1 - \frac{W_i}{2}\right) + \varepsilon_i, \\ &\text{(iii)}\ \ Y_i = W_i \left\{ e_i - 3\left(e_i - \frac{1}{2}\right)^3 \right\} + \varepsilon_i, \quad e_i = \frac{\sum_{i\neq j} E_{ij} W_j}{\sum_{i\neq j} E_{ij}}, \end{aligned} \tag{4}$$

where in each case $E(\varepsilon_i \mid W) = 0$. Here, qualitatively, Setting (i) resembles that considered by Cai et al. (2015) and Leung (2020), as discussed in Example 1; Setting (ii) exhibits a type of herd immunity where units are more sensitive to treatment when most of their neighbours are untreated; and Setting (iii) has complicated nonlinear interference effects. We then see that in Setting (i), $\tau_{\text{ADE}}(\pi)$ and $\tau_{\text{AIE}}(\pi)$ do not vary with π, but in Settings (ii) and (iii) they do and may even change signs.

This potential dependence of $\tau_{\text{ADE}}(\pi)$ and $\tau_{\text{AIE}}(\pi)$ on π is something that any practitioner using these estimands needs to be aware of. However, we believe such dependence to be largely unavoidable when seeking to define nonparametric estimands in the generality considered here. For example, when estimating indirect effects of immunization in a population where roughly 30% of units have been immunized, definitions of the type developed here could be used to support nonparametric analysis of indirect effects. Now, one should recognize that any such effects would be local to the current overall immunization rate at 30%, and would likely differ from the indirect effects we would measure at a 50% overall immunization rate. However, it seems unlikely that one could use data from a population with a 30% immunization rate to nonparametrically estimate average outcomes that might be observed at a 50% immunization rate; rather, to do so, one would need to either posit a model for how infections spread or collect different data.

SUPPLEMENTARY MATERIAL

The Supplementary Material includes additional results and proofs.

REFERENCES

ARONOW, P. M. & SAMII, C. (2017). Estimating average causal effects under general interference, with application to a social network experiment. *Ann. Appl. Statist.* **11**, 1912–47.

ARONOW, P. M., SAMII, C. & WANG, Y. (2021). Design-based inference for spatial experiments with interference. *arXiv:* 2010.13599v2.

BAKSHY, E., ROSENN, I., MARLOW, C. & ADAMIC, L. (2012). The role of social networks in information diffusion. In *Proc. 21st Int. Conf. on World Wide Web*. New York: Association for Computing Machinery, pp. 519–28.

BASSE, G. & FELLER, A. (2018). Analyzing two-stage experiments in the presence of interference. *J. Am. Statist. Assoc.* **113**, 41–55.

BASSE, G. W., FELLER, A. & TOULIS, P. (2019). Randomization tests of causal effects under interference. *Biometrika* **106**, 487–94.

BOND, R. M., FARISS, C. J., JONES, J. J., KRAMER, A. D., MARLOW, C., SETTLE, J. E. & FOWLER, J. H. (2012). A 61-million-person experiment in social influence and political mobilization. *Nature* **489**, 295–8.

CAI, J., DE JANVRY, A. & SADOULET, E. (2015). Social networks and the decision to insure. *Am. Econ. J. Appl. Econ.* **7**, 81–108.

CARNEIRO, P., HECKMAN, J. J. & VYTLACIL, E. (2010). Evaluating marginal policy changes and the average effect of treatment for individuals at the margin. *Econometrica* **78**, 377–94.

CHETTY, R. (2009). Sufficient statistics for welfare analysis: A bridge between structural and reduced-form methods. *Annu. Rev. Econ.* **1**, 451–88.

ECKLES, D., KARRER, B. & UGANDER, J. (2017). Design and analysis of experiments in networks: Reducing bias from interference. *J. Causal Infer.* **5**. DOI: 10.1515/jci-2015-0021.

HALLORAN, M. E. & STRUCHINER, C. J. (1995). Causal inference in infectious diseases. *Epidemiology* **6**, 142–51.

HECKMAN, J. J., LOCHNER, L. & TABER, C. (1998). General-equilibrium treatment effects: A study of tuition policy. *Am. Econ. Rev.* **88**, 381–6.

HUDGENS, M. G. & HALLORAN, M. E. (2008). Toward causal inference with interference. *J. Am. Statist. Assoc.* **103**, 832–42.

KARWA, V. & AIROLDI, E. M. (2018). A systematic investigation of classical causal inference strategies under misspecification due to network interference. *arXiv:* 1810.08259.

LEUNG, M. P. (2020). Treatment and spillover effects under network interference. *Rev. Econ. Statist.* **102**, 368–80.

LI, S. & WAGER, S. (2020). Random graph asymptotics for treatment effect estimation under network interference. *arXiv:* 2007.13302.

MANSKI, C. F. (2013). Identification of treatment response with social interactions. *Economet. J.* **16**, S1–23.

MIGUEL, E. & KREMER, M. (2004). Worms: Identifying impacts on education and health in the presence of treatment externalities. *Econometrica* **72**, 159–217.

MUNRO, E., WAGER, S. & XU, K. (2021). Treatment effects in market equilibrium. *arXiv:* 2109.11647.

NEYMAN, J. (1923). Sur les applications de la théorie des probabilités aux experiences agricoles: Essai des principes. *Roczniki Nauk Rolniczych* **10**, 1–51.

ROGERS, T. & FELLER, A. (2018). Reducing student absences at scale by targeting parents' misbeliefs. *Nature Hum. Behav.* **2**, 335–42.

RUBIN, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *J. Educ. Psychol.* **66**, 688–701.

SACERDOTE, B. (2001). Peer effects with random assignment: Results for Dartmouth roommates. *Quart. J. Econ.* **116**, 681–704.

SÄVJE, F. (2021). Causal inference with misspecified exposure mappings. *arXiv:* 2103.06471.

SÄVJE, F., ARONOW, P. M. & HUDGENS, M. G. (2021). Average treatment effects in the presence of unknown interference. *Ann. Statist.* **49**, 673–701.

SINCLAIR, B., MCCONNELL, M. & GREEN, D. P. (2012). Detecting spillover effects: Design and analysis of multilevel experiments. *Am. J. Polit. Sci.* **56**, 1055–69.

TCHETGEN TCHETGEN, E. J. & VANDERWEELE, T. J. (2012). On causal inference in the presence of interference. *Statist. Meth. Med. Res.* **21**, 55–75.

VANDERWEELE, T. J. & TCHETGEN TCHETGEN, E. J. (2011). Effect partitioning under interference in two-stage randomized vaccine trials. *Statist. Prob. Lett.* **81**, 861–9.

WAGER, S. & XU, K. (2021). Experimenting in equilibrium. *Manag. Sci.* **67**, 6629–7289.

[*Received on* 30 *April* 2021. *Editorial decision on* 10 *January* 2022]