Statistical Science
2024, Vol. 39, No. 3, 375–390
https://doi.org/10.1214/23-STS911

An Introduction to Proximal Causal Inference

Eric J. Tchetgen Tchetgen, Andrew Ying, Yifan Cui, Xu Shi, and Wang Miao

Abstract. A standard assumption for causal inference from observational data is that one has measured a sufficiently rich set of covariates to ensure that within covariate strata, subjects are exchangeable across observed treatment values. Skepticism about the exchangeability assumption in observational studies is often warranted because it hinges on investigators' ability to accurately measure covariates capturing all potential sources of confounding. Realistically, confounding mechanisms can rarely if ever, be learned with certainty from measured covariates. One can therefore only ever hope that covariate measurements are at best proxies of true underlying confounding mechanisms operating in an observational study, thus invalidating causal claims made on basis of standard exchangeability conditions. Causal inference from proxies is a challenging inverse problem which has to date remained unresolved. In this paper, we introduce a formal potential outcome framework for *proximal causal inference*, which while explicitly acknowledging covariate measurements as imperfect proxies of confounding mechanisms, offers an opportunity to learn about causal effects in settings where exchangeability on the basis of measured covariates fails. The proposed framework is closely related to the emerging literature on the use of proxies or negative control variables for nonparametric identification of causal effects in presence of hidden confounding bias (*Biometrika* **105** (2018) 987–993). However, while prior literature largely focused on point treatment settings, here we consider the more challenging setting of a complex longitudinal study with time-varying treatments and both measured and unmeasured time-varying confounding. Upon reviewing existing results for proximal identification in the point treatment setting, we provide new identification results for the time-varying setting, leading to the *proximal g-formula* and corresponding *proximal g-computation algorithm* for estimation. These may be viewed as generalizations of Robins' foundational g-formula and g-computation algorithm, which account explicitly for bias due to unmeasured confounding. Applications of proximal g-computation of causal effects are given for illustration in both point treatment and time-varying treatment settings.

Key words and phrases: Causality, counterfactual outcomes, proxies, confounding, negative control.

1. INTRODUCTION

A key assumption routinely made for causal inference from observational data is that one has measured a sufficiently rich set of covariates, to ensure that within covariate strata, subjects are exchangeable across observed treatment values [21, 26]. This fundamental assumption is inherently untestable empirically, without introducing a different untestable assumption, and therefore must be taken on faith even with substantial subject matter knowl-

Eric J. Tchetgen Tchetgen is Professor in the Department of Statistics and Data Science, The Wharton School, University of Pennsylvania, Philadelphia, Pennsylvania 19104, USA (e-mail: ett@wharton.upenn.edu). Andrew Ying is postdoctoral fellow in the Department of Statistics and Data Science, the Wharton School, University of Pennsylvania, Philadelphia, Pennsylvania 19104, USA. Yifan Cui is Assistant Professor in the Center for Data Science and the School of Management, Zhejiang University, Hangzhou, China. Xu Shi is Assistant Professor in the Department of Biostatistics, University of Michigan, Ann Arbor, Michigan 48109, USA. Wang Miao is Assistant Professor in the Department of Probability and Statistics, Peking University, Bejing, China.

edge at hand. For this reason, the assumption of exchangeability in observational studies is often the subject of much skepticism, mainly because it hinges on an assumed ability of the investigator to accurately measure covariates relevant to the various confounding mechanisms potentially present in a given observational study. Realistically, confounding mechanisms, that is, the set of factors that include all common causes of the treatment and outcome variables can rarely if ever, be learned and accounted for with certainty from measured covariates. Therefore, the most one can hope for in practice, is that covariate measurements are at best proxies of the true underlying confounding mechanism operating in a given observational study. Such acknowledgment invalidates any causal claim made on the basis of exchangeability.

We contribute to the literature by formally introducing a general framework for *proximal causal inference*, which while explicitly acknowledging covariate measurements as imperfect proxies of confounding mechanisms, enables one to potentially learn about causal effects in settings where exchangeability does not hold on the basis of measured covariates. This framework is closely related to the negative control framework [34], which provides sufficient conditions for nonparametric identification, leading to the *proximal g-formula* and corresponding *proximal g-computation algorithm* for estimation in point exposure studies. These may be viewed as generalizations of Robins' foundational g-formula and g-computation algorithm, which account explicitly for bias due to unmeasured confounding. In addition to reviewing existing identification results in the point exposure setting, a key contribution of this paper is to extend the framework to longitudinal studies with time-varying treatment, to account for both measured and unmeasured time-varying confounding.

As all formal methods for causal inference, the proposed proximal approach relies on assumptions that are not testable empirically without a different assumption; nevertheless, as we argue next, we view the required identifying assumptions as easily interpretable and potentially easier to reason about on subject matter grounds than exchangeability. Mainly, proximal causal inference requires that the analyst can correctly classify proxies into three bucket types: a. variables which are common causes of the treatment and outcome variables; b. treatment confounding proxies versus c. outcome confounding proxies. A proxy of type b is a potential (not necessarily a) correlate of the treatment which is related with the outcome conditional on treatment only through an unmeasured common cause for which the variable is a proxy; while a proxy of type c is a potential (not necessarily a) cause of the outcome which is related with the treatment only through an unmeasured common cause for which the variable is a proxy. Proxies that are neither directly related to the treatment or outcome variable other than through the unmeasured confounder may belong to either bucket type b or c.

In order to ground ideas, we briefly describe the proposed proximal approach in the context of a point exposure A, outcome Y, and unmeasured confounder U; then suppose that one can correctly select a treatment proxy Z and an outcome proxy W such that the simple structural linear model given below holds:

$$\begin{aligned} E(Y|A,Z,X,U) &= \beta_0 + \beta_a A + \beta_u U + \beta_x' X, \\ E(W|A,Z,X,U) &= \eta_0 + \eta_u U + \eta_x' X, \end{aligned} \tag{1}$$

where X are all other observed covariates, and validity of proxies is encoded by the fact that the right-hand side of the first equation does not depend on Z, the right-hand side of the second equation does not depend on A and Z; and W is U relevant in the sense that $\eta_u \neq 0$. This system of linear structural models is compatible with the causal diagram in Figure 2(b) of the next section, and may also be viewed as a marginal structural model for the causal diagram in Figure 1(b) whereby W is marginalized over in the outcome model. The causal parameter of interest is $\beta_a = E(Y_{a+1} - Y_a|U,X)$ corresponding to the average outcome difference if one were to intervene to increase the treatment by one unit upon conditioning on covariates (U,X) a sufficient confounding adjustment set; that is exchangeability holds conditional on (U,X). It is then straightforward to show that

$$\begin{aligned} E(Y|A,Z,X) &= \beta_0 + \beta_a A + \beta_u E(U|A,Z,X) \\ &\quad + \beta_x' X, \\ E(W|A,Z,X) &= \eta_0 + \eta_u E(U|A,Z,X) + \eta_x' X, \end{aligned} \tag{2}$$

$$\text{so that} \quad E(Y|A,Z,X) = \beta_0^* + \beta_a A + \beta_u^* E(W|A,Z,X) + \beta_x^{*\prime} X,$$

where $\beta_0^* = \beta_0 - \frac{\beta_u \eta_0}{\eta_u}$, $\beta_u^* = \frac{\beta_u}{\eta_u}$, $\beta_x^* = \beta_x - \frac{\beta_u \eta_x}{\eta_u}$.

Let $\widehat{W}$ denote an (asymptotically) unbiased estimator of $E(W|A,Z,X)$, then equation (2) suggests that provided that $E(U|A,Z,X)$ depends on Z, the least squares linear regression of Y on $(A,X,\widehat{W})$ recovers a slope coefficient for A, $\widehat{\beta}_a$ that is consistent for the causal parameter β_a. In contrast, either removing $\widehat{W}$ from the regression model, or replacing it with either W, or Z, or (W,Z) will generally yield a biased estimate of β_a given that exchangeability does not hold either conditional on X, on (X,W), or on (X,W,Z). As further adjusting for $\widehat{W}$ debiases the least squares estimator of β_a conditional on X, we shall refer to $\widehat{W}$ as a *proximal control variable*. The system of linear structural equations considered above is overly restrictive, assuming linearity and no interactions; as we demonstrate in this paper, these assumptions are not strictly necessary and can be relaxed considerably so that nonparametric identification remains possible under certain conditions.

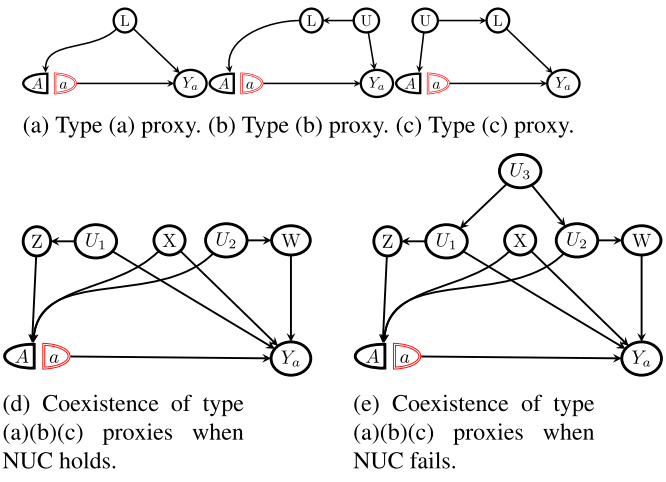
The paper is organized as follows. Notation and formal definitions used throughout the paper are given in Section 2, where we also review standard identification of causal effects assuming exchangeability. In Section 3, we introduce the proximal causal inference framework and formally define our three types of proxies. We provide a number of examples of proxies in observational studies and also draw connections with literature on negative controls. Drawing heavily on Miao, Geng and Tchetgen Tchetgen [34], we describe nonparametric identification conditions for proximal causal inference, in point exposure settings by the *proximal g-formula*, a generalization of Robins' foundational g-formula which accounts for confounding bias due to unmeasured factors. Subsequently, we generalize the framework to longitudinal studies involving time-varying treatments and time-varying confounding in Section 4, where we establish nonparametric proximal identification of the joint causal effects of time-varying treatments. These results are entirely new to the causal inference literature. For estimation, a *proximal g-computation algorithm* is then introduced in Section 5. As we show, equation (2) can be recovered as a special case of proximal g-computation algorithm under certain linearity assumptions. Estimation and inference for both point treatment and time-varying treatment settings is then discussed in some level of detail. Applications of proximal causal inference are given to illustrate the methodology in both point exposure and time-varying exposure settings in Section 6. The paper concludes with brief final remarks in Section 7.

2. IDENTIFICATION UNDER STANDARD EXCHANGEABILITY

Suppose one has observed i.i.d. samples on (A, L, Y) where as before A denotes a treatment of interest, Y is an outcome of interest, and let L denote a set of measured covariates. Let Y_a denote the potential outcome had, possibly contrary to fact a person received treatment $A = a$. Throughout, we make the standard consistency assumption linking observed and potential outcomes

$$Y = Y_A, \tag{3}$$

almost surely. We aim to identify a population average causal effect, corresponding to a contrast of counterfactual averages $\beta(a) = E(Y_a)$ for different values of a. For instance, in case of binary treatment, one might be interested in the average treatment effect measured on the additive scale $\beta(1) - \beta(0) = E(Y_1) - E(Y_0)$; in case of binary outcome, one might also be interested in the average treatment effect on the multiplicative scale $\beta(1)/\beta(0) = \Pr(Y_1 = 1)/\Pr(Y_0 = 1)$. In all cases, whether A is binary, polytomous or continuous, learning about causal effects on any given scale involves learning about the potential outcome mean $\beta(a)$, which we aim to identify from the observed sample. A common identification strategy in observational studies is that of exchangeability [21, 26, 43] or no unmeasured confounding (NUC) condition on the basis of measured covariates:



(a) Type (a) proxy. (b) Type (b) proxy. (c) Type (c) proxy.

(d) Coexistence of type (a)(b)(c) proxies when NUC holds.

(e) Coexistence of type (a)(b)(c) proxies when NUC fails.

FIG. 1. *Single world intervention graphs illustrating treatment and outcome confounding proxies.*

$$Y_a \perp\!\!\!\perp A|L, \tag{4}$$

where $\perp\!\!\!\perp$ denotes independence, together with positivity condition that

$$f(A = a|L) > 0, \tag{5}$$

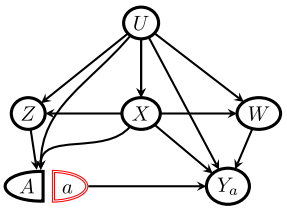
almost surely, where $f(D|G)$ denotes the conditional density or probability mass function of D given G. Assumption (4) is sometimes interpreted as stating that L includes all common causes of A and Y; an assumption represented in single world intervention graph (SWIG) in Figure 1(a) in which L is of Type a.

Under assumptions (3)–(5), it is well known that

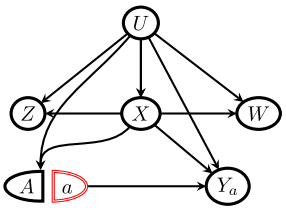
$$\beta(a) = \sum_l E(Y|a, l) f(l), \tag{6}$$

a formula most commonly known in the field of epidemiology as the g-formula [21], a name associated with the work of James Robins which we shall adopt in this paper. It is also known as the back-door adjustment formula because L satisfies the back-door criterion [40].

It is interesting to consider alternative data generating mechanisms under which assumption (4) holds, illustrated in Figures 1(b) and 1(c), with the first of Type b where L includes all causes of A that share an unmeasured common cause U (and therefore are associated) with Y; while the second is of Type c where L includes all causes of Y that share an unmeasured common cause U (and therefore are associated) with A. These three types may coexist, as displayed in Figure 1(d) in which L has been decomposed into three types of measured covariates $L = (X, W, Z)$, such that X are measured covariates of Type a, Z are measured covariates of Type b, while W are



(a) Coexistence of type (a)(b)(c) proxies when NUC fails.



(b) Coexistence of type (a)(b)(c) proxies when NUC fails.

FIG. 2. *Single world intervention graphs with endogenous point exposure and proxies.*

measured covariates of Type c. At any rate, all settings represented in Figure 1(a)–Figure 1(d) illustrate possible data generating mechanisms under which exchangeability assumption (4) holds, without necessarily requiring that the analyst identifies which bucket type each covariate in $L = (X, Z, W)$ belongs to. Importantly, all these settings rule out the presence of an unmeasured common cause of A and Y conditional on L, therefore ruling out unmeasured confounding. Note that in order for exchangeability to hold in Figure 1(d), it must be that as encoded in the SWIG, unmeasured variables U_1 and U_2 do not introduce a new backdoor path between A and Y as illustrated in Figure 1(e), the unblocked backdoor path $A - U_2 - U_3 - U_1 - Y$ would invalidate assumption (4). As we show in the next section, it is sometimes possible to relax this last assumption and therefore exchangeability condition (4) while preserving identification of $\beta(a)$ despite the presence of unmeasured confounding, provided that one can correctly identify which bucket type each measured covariate falls into.

3. PROXIMAL IDENTIFICATION IN POINT EXPOSURE STUDIES

Next, consider Figures 2(a) and 2(b) which depict settings in which exchangeability condition (4) fails, despite having measured covariates $L = (X, Z, W)$, owing to the presence of an unmeasured common cause U of A and Y. The SWIG in Figure 2(a) may be viewed as a generalization of Figure 1(e) by letting $U = (U_1, U_2, U_3)$. In the following, we propose to replace the untestable assumption (4) with an assumption that the analyst has correctly identified variables in bucket Types b and c. Formally, the SWIG in Figure 2(a) implies that

$$(W, Y_a) \perp\!\!\!\perp (A, Z)|U, X. \tag{7}$$

The conditional independence in above display formally encodes the assumption that adjusting for (X, U) would in principle suffice to identify the joint causal effect of $(A$ on Y and W, respectively; likewise, conditioning on (X, U, A) d-separates Z from (Y, W). This is a latent exchangeability assumption which is reasonable as long as there exists a U sufficiently enriched to include all common causes of (A, Z) and (Y, W) not included in X. As it is not required that U be observed, the assumption will generally hold even in observational studies.

We formally refer to Type b variables Z as treatment confounding proxies and Type c variables W as outcome confounding proxies, provided that they satisfy (7). It is important to note that Z may be independent of A given (X, U) (in which case the edge between Z and A can be omitted as displayed in Figure 2(b)), and likewise W may not be associated with Y given (X, U) (in which case the edge between W and Y can be omitted as displayed in Figure 2(b)), however as long as both are U-relevant (i.e., being sufficiently associated with U, conditional on (A, X), and satisfy (7), they are considered valid proxies for our purposes and may be allocated as type b or c at the analyst's discretion. We note that (7) does not require conceptualizing a potential intervention on covariates Z. Connection of treatment confounding proxies and instrumental variables is drawn in Remark 3 and Table A.1 in the Supplementary Material.

To summarize, we formally make the following assumptions in proximal causal inference for point exposure studies.

ASSUMPTION 1 (Consistency).

$$Y = Y_A, \tag{8}$$

almost surely.

ASSUMPTION 2 (Positivity).

$$f(A = a|X, U) > 0, \tag{9}$$

almost surely for $a = 0, 1$.

ASSUMPTION 3 (Point Proxies and Latent Exchangeability).

$$(W, Y_a) \perp\!\!\!\perp (A, Z)|U, X. \tag{10}$$

REMARK 1. A set of weaker assumptions implied by Assumptions 1–3 is also sufficient for proximal identification. This includes Assumptions 1–2, latent exchangeability

$$Y_a \perp\!\!\!\perp A|U, X, \tag{11}$$

and conditional independence on the observed variables,

$$W \perp\!\!\!\perp (A, Z)|U, X, \tag{12}$$

$$Y \perp\!\!\!\perp Z|A, U, X. \tag{13}$$

3.1 Examples of Proxies

Examples of proxies of type b and c abound in observational studies. For instance, in an observational study

evaluating the effects of a treatment on disease progression, one is typically concerned that patients either self-select or are selected by their physician to take the treatment based on prognostic factors for the outcome; therefore there may be two distinct processes contributing to a subject's propensity to be treated. In an effort to account for these sources of confounding, a diligent investigator would endeavor to record lab measurements and other clinically relevant covariate data available to the physician and the patient when considering treatment options. For instance, it is customary in evaluating the effectiveness of HIV anti-retroviral therapy, to adjust for CD4 count measurement as a potential source of confounding by indication [24]; this is because whenever available to the prescribing physician, CD4 count measurement is invariably used to decide (at least prior to the advent of universal test and treat) whether a patient should be given ART, that is, the probability of treatment initiation generally decreases with a patient's increasing CD4 count. However, as an error-prone snapshot measurement of the evolving state of the patient's underlying immune system, CD4 count measurement is unlikely to be a direct cause of disease progression but rather a proxy of the actual state of patient's immune system, the actual cause of disease progression. It is therefore more accurate to conceive of baseline CD4 count measurement as an imperfect proxy of immune system status. In addition, two patients with similar CD4 count measurements seen by different physicians may differ in terms of treatment decisions depending on the patient's own health seeking behavior, as well as differences in physician's clinical training and experience, and overall prescription preferences; which are all factors potentially predictive of patient disease progression regardless of treatment. Because such factors are notoriously difficult to measure accurately, they are likely to induce residual confounding, even after adjusting for baseline CD4 count.

A proxy of type c might include baseline covariate measurements assessing a patient's mental and physical comorbidities including those measured using a validated questionnaire. For instance, it is well known that in addition to the immune system, HIV affects the nervous system and the brain producing neurological sequelae, often resulting in forgetfulness and cognitive problems [23, 55]. These problems can compromise medication adherence, interfere with essential activities of daily living such as driving and managing finances, increase dependency, and decrease quality of life. Several cognitive functioning screening tools exist to objectively measure cognition, with the gold standard being the Mini-Mental State Examination (MMSE) [17, 52], which is a 30-point questionnaire that is used extensively in clinical and research settings. A score of 24 or less is generally used as a cut-off to indicate possible mild cognitive impairment or early stage dementia [17] although the cut-off can vary according to the education level of the individual [51]. Although widely used as a measure of cognition, MMSE is at best an imperfect proxy of a patient's baseline state of cognitive impairment, which may in turn influence both a patient's willingness to initiate and adhere to ART, and the patient's disease progression at follow-up. Thus, in evaluating the causal effect of ART on disease progression such as say cognitive decline, baseline MMSE measurements can be seen as a proxy of type c for the underlying confounding mechanism corresponding to the patient's underlying state of cognitive impairment at baseline. These are but two motivating examples of confounding proxies in analysis of the causal effects of ART on HIV infection related disease progression from an observational study. Aside for these proxies, there may also be factors that can accurately be described as true common causes of treatment and outcome processes; these variables which we have referred to as of type a may in fact include age, gender and years of education depending on the context. Thus, rather than as current practice dictates, assuming that adjusting for baseline covariates, exchangeability can be attained, our proposed proximal framework requires the investigator correctly classifies covariates that belong in bucket types a, b and c without necessarily the need for exchangeability to hold conditional on such proxies.

REMARK 2. In prior work, variables of Types b and c satisfying (10) have been called negative control exposure and negative control outcome variables, referring to negative control variables an investigator would need to supplement her observational study with, such that (10) would be satisfied. In this paper, we prefer the proxy terminology to the negative control nomenclature, to highlight the key observation, that often, covariates measured in an observational study in an effort to control for confounding, may not be sufficient to fulfill exchangeability, but nevertheless can potentially be partitioned into proxies satisfying negative control condition (10). This observation, therefore, alleviates the need to supplement one's observational study design by collecting additional data on potential negative control variables, although variables of Type b and c may be enriched with appropriately selected negative control auxiliary variables when available. Our framework of leveraging proxies of unmeasured confounders is also related to the literature on adjustment of confounders measured with error [32, 39].

REMARK 3. It is further important to note that while we have taken Figure 2 as canonical representations of proxy variables of Types b and c, several alternative DAGs might be compatible with (10), as illustrated in the Supplemental Material [50], Table A.1. Interestingly, the DAG given in first row and first column of Table A.1 of the appendix establishes that an instrumental variable (IV)

for the causal effect of A on Y may be included in Type a bucket provided that it is also a valid IV for W. In fact, even an invalid instrumental variable which fails to satisfy the IV independence assumption [25, 54] may also be included in bucket Type b as (10) are satisfied [37].

REMARK 4. Additionally, one should note that similar to exchangeability condition (4), (10) is not empirically testable as they presume certain null causal effects and involve conditional independence statements given the unmeasured variable U. Interestingly, the condition that (Z, A) are conditionally independent of W, is a given in instances where W and Z are contemporaneous (and therefore cannot cause each other), and are pre-treatment covariates. The treatment can therefore not have a causal effect on W as the future cannot cause the past. It is sometimes reasonable to include post-outcome variables in Z so that Assumption 3 holds by design, again due to temporal ordering. However, in order to satisfy (10), there must be no unblocked causal pathway between (Y, W) and Z conditional on U, X and A. Likewise, it is sometimes possible to include pre-treatment measurements of the outcome in view as potential negative control outcome and therefore to include them in W, provided that they satisfy (10) and therefore do not have a direct effect on treatment and negative control exposure variables [37, 36, 48]. An example was provided in Miao, Shi and Tchetgen Tchetgen [36], Miao and Tchetgen Tchetgen [37] in the context of studying the causal effect of air pollution on mortality or elderly hospitalization using time series data, in which case air pollution measurement post-hospitalization may be a reasonable choice of negative control exposure to include in Z, and hospitalization measurement pre-air pollution may likewise be a reasonable negative control outcome to include in W [36, 37].

The aforementioned connection to negative control literature is instrumental in determining sufficient conditions for nonparametric identification of $\beta(a)$ by leveraging identification results recently obtained by Miao, Geng and Tchetgen Tchetgen [34]. We summarize their results below, provide intuition for the results and refer the interested reader to their manuscript for a careful treatment of mathematical conditions underpinning the approach. In Section 4, we extend their results to the time-varying setting, which to our knowledge is new to the literature, all proofs can be found in the Appendix.

Let $h(a, x, w)$ denote a solution to the equation

$$E(Y|a, z, x) = \sum_w h(a, x, w) f(w|a, x, z), \tag{14}$$

where in slight abuse of notation $\sum$ denotes an integral in case of continuous w. Next, suppose that the following conditions hold, for any square-integrable function $v(\cdot)$ and

$$\begin{aligned} &E\{v(U)|Z, A, X\} = 0 \quad \text{almost surely} \\ &\text{if and only if} \quad v(U) = 0 \quad \text{almost surely.} \end{aligned} \tag{15}$$

This condition is formally referred to as a completeness condition which accommodates both categorical and continuous confounders. Completeness is a technical condition taught in core courses in theory of statistical inference and plays an important role in foundational theorems such as Lehmann–Scheffé theorem and Basu's theorem [7]. More recently, completeness assumption has played a key role in major identification results for a variety of nonparametric and semiparametric models in the econometrics literature, including nonparametric models with instrumental variables [38, 1, 11, 22, 4, 15, 27, 9], measurement error models [28, 2, 6, 10], dynamic models [29, 47], panel data models [19], and auction models. Therefore, our work also connects to this rich legacy in which the completeness assumption has been invoked quite widely.

Intuitively, one may interpret completeness as a basic requirement that the set of proxies must have sufficient variability relative to variability of U. This requirement can readily be described in the case of categorical U, Z and W, with number of categories d_u, d_z and d_w respectively. In this case, completeness requires that

$$\min(d_z, d_w) \geq d_u, \tag{16}$$

which states that Z and W must each have at least as many categories as U. Intuitively, condition (16) states that proximal causal inference can potentially account for unmeasured confounding in the categorical case as long as the number of categories of U is no larger than that of either proxies Z and W [45]. Failure of this assumption, say if Z is coarser than U will generally imply lack of identification, without a different assumption. Interestingly, Chen et al. [8] and Andrews [3] showed that if Z and U are continuously distributed and the dimension of Z is larger than that of U, then under relatively mild regularity conditions, the completeness condition holds generically in the sense that the set of distributions for which completeness fails has a property similar to being essentially Lebesgue measure zero. More formally, they establish that distributions for which a completeness condition fails can approximate distributions for which it holds arbitrarily well in the total variation distance sense. Thus, while completeness conditions may themselves not be directly testable, one may argue as in Canay, Santos and Shaikh [5] that they are commonly satisfied. We further refer to Andrews [3], Chen et al. [8], and Section 2 of the Supplementary Material of Miao et al. [35] for a thorough review and several examples illustrating completeness. Also see Ying et al. [57] who conducted empirical studies assessing sensitivity of proximal causal methods we introduce below, when completeness fails to hold exactly. To avoid such failure, completeness provides a rationale for measuring to the extent possible, a rich set of baseline characteristics in observational studies as a potential strategy for mitigating unmeasured confounding

via the proximal approach we now describe. Clearly, such strategy should be balanced against the efficiency loss one might incur by inadvertently including irrelevant proxies, and the potential need for larger samples to accurately estimate moderate to high dimensional nuisance function. Miao, Geng and Tchetgen Tchetgen [34] established that under Assumptions 1, 2, 3, and conditions (14), (15) are satisfied, the counterfactual mean $\beta(a)$ can be identified nonparametrically by the formula

$$\beta(a) = \sum_{w,x} h(a,x,w) f(w,x). \tag{17}$$

We refer to equation (17) as the *proximal g-formula*, and to $h(a,x,w)$ as an outcome confounding bridge function Miao, Shi and Tchetgen Tchetgen [37]. A few key observations are in order. First, equation (14) defines a so-called inverse problem formally known as a Fredholm integral equation of the first kind. Formal conditions for existence of a solution of such an equation are well established in functional analysis in mathematical literature, but due to their technical nature are beyond the scope of the current paper, though it is worth mentioning that existence of a solution requires the following additional completeness condition: for any square-integrable function $v(\cdot)$ and

$$\begin{aligned} &E\{v(Z)|W,A,X\} = 0 \quad \text{almost surely} \\ &\text{if and only if} \quad v(Z) = 0 \quad \text{almost surely}, \end{aligned} \tag{18}$$

an assumption that cannot hold unless Z and W are U-relevant; otherwise Z and W would be conditionally independent and the first equality above would hold for all $v(Z) \neq 0$ with mean zero given (A,X). In the categorical case, as established in Shi et al. [45] condition (16) along with a rank condition for a certain matrix defined in terms of the conditional distribution of W given (Z,A,X) suffices for equation (14) to admit a solution [45]. It is important to note that $h(a,x,w)$ satisfying (14) need not be unique, any solution to this equation yields the same value of the proximal g-formula. We also note that by latent exchangeability, $\beta(a) = \sum_{u,x} E(Y|a,u,x) f(u,x) = \sum_{w,x} h(a,x,w) f(w,x)$; and as shown in Miao, Geng and Tchetgen Tchetgen [34],

$$E(Y_a|u,x) = \sum_{w} h(a,x,w) f(w|u,x), \tag{19}$$

which highlights the inverse-problem nature of the task accomplished by proximal g-formula, which is to determine a function h that satisfies this equality without explicitly modeling or estimating the latent factor U. A remarkable feature of proximal causal inference is that accounting for U without either measuring U directly or estimating its distribution can be accomplished provided that the set of proxies though imperfect, is sufficiently rich so that the inverse-problem admits a solution in a model-free framework. It is also worth noting that (19) readily implies the following identifying formulae $E(Y_a|z,x) = \sum_w h(a,x,w) f(w|z,x)$ and $E(Y_a|x) = \sum_w h(a,x,w) f(w|x)$. Intuition about conditions under which a unique solution to equation (14) might exist can be gained in the simple case of binary A, W, Z, whereby it is straightforward to show that the unique solution to (14) is

$$\begin{aligned} h(a,x,w) &= E(Y|a,z,x) \\ &\quad + g(a,x)\big[w - \Pr(W=1|a,z,x)\big], \end{aligned} \tag{20}$$

where

$$g(a,x) = \frac{E(Y|a,z=1,x) - E(Y|a,z=0,x)}{\Pr(W=1|a,z=1,x) - \Pr(W=1|a,z=0,x)}.$$

Importantly, note that although the right-hand side to equation (20) appears to depend on z, the left-hand side indicates that it does not, which is readily verified with some algebra. In order for h to be finite, one requires that $\Pr(W=1|a,z=1,x) - \Pr(W=1|a,z=0,x) \neq 0$; that is W must be associated with Z conditional on (A,X), a condition that one would expect to hold to the extent that W and Z are strong proxies of U, thus further highlighting the importance of selecting strong potential proxies. In the binary case, $\beta(a)$ takes the closed form

$$\begin{aligned} \beta(a) = E_X\big\{&E(Y|a,Z,X) - g(a,X)\big[\Pr(W=1|a,X) \\ &- \Pr(W=1|a,Z,X)\big]\big\}. \end{aligned}$$

A generalization of the above closed-form expression for proximal g-formula with categorical variables is given in [45] for the average causal effect $\beta(1) - \beta(0)$ of a binary treatment on the additive scale. Unfortunately, unlike the g-formula, the proximal g-formula is not always available in closed-form and requires solving equation (14) numerically, which might be computationally intensive and unstable due to its potential to be empirically ill-posed. Ill-posedness in this case refers to the fact that small amount of uncertainty in estimating the left hand side of (14) empirically can often induce excessive uncertainty in obtaining a solution to the equation. Such ill-posedness is typically addressed by some form of regularization of the integral equation. Below, we describe a simple statistical modeling approach analogous to g-computation, which sidesteps this difficulty by automatically generating stable solutions to the equation under correct model specification. Revisiting the motivating example given in the introduction, one may readily verify that the structural equations (1) imply that there exists coefficient $\eta = (\eta_0, \eta_a, \eta_x', \eta_W)$ such that

$$\begin{aligned} &h(A,X,W;\eta) = \eta_0 + \eta_a A + \eta_x' X + \eta_W W, \quad \text{and} \\ &E\big(Y - h(A,X,W;\eta)|A,Z,X\big) = 0, \end{aligned}$$

so that equation (14) is satisfied. By then applying the proximal g-formula, one recovers

$$\beta_a = \eta_a = E\{h(A=1, X, W; \eta) - h(A=0, X, W; \eta)\},$$

and identifies the causal effect parameter.

It is interesting to compare proximal g-formula to standard g-formula [21]. In this vein, suppose that $L = (X, W)$ suffices for exchangeability condition (4), so that $Z = \varnothing$; then, proximal g-formula reduces to the standard g-formula with $h(a, x, w) = E(Y|a, x, w)$ a stable solution to (14) given that

$$\begin{aligned} E(Y|a, x) &= \sum_w h(a, x, w) f(w|a, x) \\ &= \sum_w E(Y|a, x, w) f(w|a, x), \\ \beta(a) &= \sum_{w,x} h(a, x, w) f(w, x) \\ &= \sum_{w,x} E(Y|a, x, w) f(w, x). \end{aligned}$$

From this perspective, exchangeability may be viewed as a form of regularization of equation (14) which automatically yields a unique stable solution to the integral equation.

REMARK 5. We note that Miao, Shi and Tchetgen Tchetgen [37] considered alternative identifying conditions in that instead of taking equation (14) as starting point, they a priori assume that there exist a bridge function $h(w, a, x)$ such that $E(Y_a|u, x) = \sum_w h(a, x, w) f(w|u, x)$; in addition, they replace completeness condition (15) which is not subject to an empirical test, with the testable completeness condition that $E\{v(w)|z, a, x\} = 0$ for all z, a and x if and only if $v(w) = 0$; then they establish that such function h must solve equation (14).

REMARK 6. We also briefly note that under the causal graph in Figure 2(b) one could in principle verify whether the SWIG is misspecified by evaluating the proximal g-formula upon permuting variables allocated to W and Z respectively. That is, Z and W are symmetric under Figure 2(b) and can be switched. Under the null hypothesis that the graph is correctly specified, we would then expect that the proximal g-formula would remain invariant to the choice of W and Z. A simple test of invariance of the estimated causal effect over a pair of permutations of Z and W can then be obtained by standardizing the difference in the estimated causal effects by a corresponding estimate of the standard error of the estimated difference. We expect such a test statistic to have reasonable power against important departures from the graphical model of Figure 2(b) provided that Z and W are relevant for U.

4. PROXIMAL IDENTIFICATION IN COMPLEX LONGITUDINAL STUDIES

We now consider proximal identification of causal effects in complex longitudinal studies. In order to ground ideas and simplify the exposition, we focus primarily on a special case of a longitudinal study with two follow-up times and briefly review identification under a longitudinal version of exchangeability. Thus, suppose that one has observed time-varying treatment and covariate data $\{L(j), A(j)\}$ at follow up visits $j = 0, 1$ of a longitudinal study. Let Y denote the outcome of interest measured at the end of follow-up $j = 2$. We assume that recorded data on the treatment and prognostic factors do not change except at these times, moreover, $L(j)$ temporally precedes $A(j)$. We use overbars to denote the history of that variable up to end of follow-up; for example, $\overline{L} = \{L(0), L(1)\}$. Let $Y_{\overline{a}} = Y_{a(0),a(1)}$ denote the potential outcome had possibly contrary to fact, a subject followed treatment regime $\overline{A} = \overline{a}$. Our aim is to identify the potential outcome mean $\beta(\overline{a}) = E(Y_{\overline{a}})$. To do so, three standard assumptions are typically invoked. The first entails a longitudinal version of consistency:

$$Y = Y_{\overline{A}},$$

almost surely, linking counterfactual outcomes $\{Y_{\overline{a}} : \overline{a}\}$ to observed variables $(Y, \overline{A})$. The next assumption is that there are no unmeasured confounders for the effect of $A(j)$ on Y, that is, for all treatment histories $\overline{a}$,

$$\begin{aligned} &Y_{\overline{a}} \perp\!\!\!\perp A(0)|L(0), \quad \text{and} \\ &Y_{\overline{a}} \perp\!\!\!\perp A(1)|A(0) = a(0), \overline{L}. \end{aligned} \tag{21}$$

This assumption which generalizes exchangeability to the longitudinal setting is also known as the sequential randomization assumption (SRA) [41]. It states that conditional on treatment history and the history of all recorded covariates up to j, treatment at j is essentially randomized by nature and thus must be independent of the counterfactual random variable $Y_{\overline{a}}$ [26, 42].

We finally assume that the following positivity assumption holds. For all $a(j)$ in the support of $A(j)$ if

$$\begin{aligned} &f\big(\overline{L}(j), \overline{A}(j-1)\big) > 0 \quad \text{then} \\ &f\big(a(j)|\overline{L}(j), \overline{A}(j-1)\big) > 0, \quad j = 0, 1 \end{aligned}$$

with $A(-1) \equiv 0$, which essentially states that if any set of subjects at time j have the opportunity of continuing on a treatment regime $\overline{a}$ under consideration, at least some will take that opportunity. Robins established that under these assumptions, the counterfactual mean $\beta(\overline{a})$ is given by the longitudinal g-formula [26, 41]:

$$\beta(\overline{a}) = \sum_{\overline{l}} E(Y|\overline{a}, \overline{l}) \prod_{j=1}^{2} f\big(l(j)|\overline{l}(j-1), \overline{a}(j-1)\big).$$

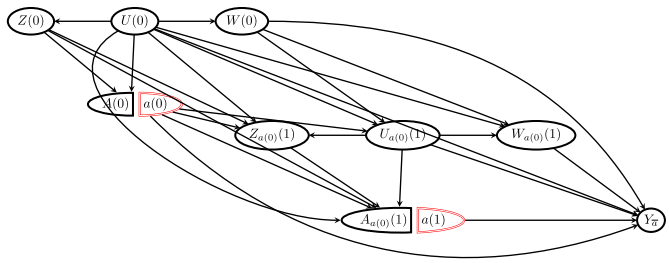


FIG. 3. *A single world intervention graph (SWIG) with endogenous time varying treatments and proxies.*

As argued in the introduction, in an observational study, the sequential randomization assumption cannot be guaranteed to hold, and it is not subject to empirical test, even when good efforts are made to collect data on crucial covariates. As before, our aim is to relax sequential exchangeability/SRA by explicitly incorporating measured covariates as proxies of underlying confounding mechanisms longitudinally. We assume the following.

ASSUMPTION 4 (Consistency).

$$Y = Y_{\overline{A}}, \tag{22}$$

almost surely.

ASSUMPTION 5 (Sequential Positivity).

$$f\big(A(1) = a(1)|A(0), \overline{X}, \overline{U}\big) > 0, \tag{23}$$

$$f\big(A(0) = a(0)|X(0), U(0)\big) > 0, \tag{24}$$

almost surely for $a(1) = 0, 1$ and $a(0) = 0, 1$.

ASSUMPTION 6 (Sequential Proxies).

$$\big(Z(0), A(0)\big) \perp\!\!\!\perp \big(\overline{W}(0), Y_{\overline{a}}\big)|\overline{U}(0), X(0), \tag{25}$$

$$\begin{aligned}&\big(\overline{Z}(1), A(1)\big) \perp\!\!\!\perp \big(\overline{W}(1), Y_{\overline{a}}\big)|\overline{U}(1), A(0)\\ &\quad = a(0), \overline{X}(1).\end{aligned} \tag{26}$$

These conditions are a longitudinal generalization of (10). Figure 3 illustrates a possible data generating mechanism in which conditions (25) and (26) hold, where to simplify the figure certain edges are shaded and we have suppressed observed time-varying covariates $\overline{X}$ which structurally follow the same relationship as $\overline{U}$ with other variables. Additionally, for identification we require the following longitudinal generalization of completeness condition (15), which state that for any square-integrable function $v(\cdot)$:

$$\begin{aligned}E\big(v(\overline{U})|\overline{A}, \overline{Z}, \overline{X}\big) = 0 \quad &\text{a.s.}\\ \text{if and only if} \quad v(\overline{U}) = 0 \quad &\text{a.s.},\end{aligned} \tag{27}$$

$$\begin{aligned}E\big(v\big(U(0)\big)|A(0), Z(0), X(0)\big) = 0 \quad &\text{a.s.}\\ \text{if and only if} \quad v\big(U(0)\big) = 0 \quad &\text{a.s.}\end{aligned} \tag{28}$$

Finally, extending assumption (14) to longitudinal setting, we suppose that there exist functions $H_1(\overline{a}) = h_1(\overline{W}, \overline{a}, \overline{X})$ and $H_0(\overline{a}) = h_0(W(0), \overline{a}, X(0))$ that solve the equations

$$E(Y|\overline{z}, \overline{a}, \overline{x}) = E\big(H_1(\overline{a})|\overline{z}, \overline{a}, \overline{x}\big), \tag{29}$$

$$E\big(H_1(\overline{a})|z(0), \overline{a}, x(0)\big) = E\big(H_0(\overline{a})|z(0), \overline{a}, x(0)\big). \tag{30}$$

Confounding bridge equation (29) is exactly equivalent to equation (14) with $(\overline{W}, \overline{Z}, \{\overline{X}, A_0\})$ replacing (W, Z, X), and $A(1)$ replacing A. As shown in the result below, this assumption yields identification of $E(Y_{a(1)}|a(0), \overline{x})$, while the second assumption (30) which does not have a point exposure analog, yields identification of $E(Y_{\overline{a}}|x(0))$. In fact, we have the following result.

THEOREM 4.1. *Under Assumptions 4–6 and conditions (27)–(30) are satisfied, then we have that*

$$\begin{aligned}E\big(Y_{a(1)}|a(0), \overline{u}, \overline{x}\big) &= E\big(H_1(\overline{a})|a(0), \overline{u}, \overline{x}\big),\\ E\big(Y_{\overline{a}}|\, u(0), x(0)\big) &= E\big\{H_0(\overline{a})|u(0), x(0)\big\},\end{aligned}$$

and

$$\begin{aligned}E\big(Y_{a(1)}|a(0), \overline{x}\big) &= E\big(H_1(\overline{a})|a(0), \overline{x}\big),\\ E\big(Y_{\overline{a}}|x(0)\big) &= E\big\{H_0(\overline{a})|x(0)\big\},\\ \beta(\overline{a}) &= E(Y_{\overline{a}})\\ &= E\big\{H_0(\overline{a})\big\} = E\big\{h_0\big(W(0), \overline{a}, X(0)\big)\big\}.\end{aligned}$$

Theorem 4.1 can be extended to a longitudinal study of follow-up length $J > 2$ as shown in the Appendix; the proof of which implies Theorem 4.1 as a special case. As in the point treatment setting, $H_1(\overline{a})$ and $H_0(\overline{a})$ need not be uniquely identified in order for $E(Y_{a(1)}|a(0), \overline{x})$, $E(Y_{\overline{a}}|x(0))$ and $\beta(\overline{a})$ to be uniquely identified. Similar to point exposure case, The theorem also implies identification of the conditional average treatment effect functions

$$\begin{aligned}E\big(Y_{a(1)}|a(0), \overline{x}, \overline{z}\big) &= E\big(H_1(\overline{a})|a(0), \overline{x}, \overline{z}\big),\\ E\big(Y_{\overline{a}}|x(0), z(0)\big) &= E\big\{H_0(\overline{a})|x(0), z(0)\big\}.\end{aligned}$$

5. PROXIMAL G-COMPUTATION

In this section, we describe a practical approach for estimating the proximal g-formula. We first describe the approach in the point treatment case before extending it to the case of time-varying treatment. Thus, suppose that one has observed an i.i.d. sample of size n on $(A, L = (X, Z, W))$. It is then convenient to directly specify a parametric model for the outcome bridge function,

$$h(W, A, X) = h(W, A, X; \eta),$$

with unknown parameter η; and for the joint law

$$f(L, A) = f(L, A; \theta),$$

with unknown parameter θ. Note that by (14), together, these models entail a parametric model for

$$\begin{aligned}\mu(A; \eta, \theta) &= E(Y|A, X, Z; \eta, \theta)\\ &= \sum_w f(w|Z, A, X; \theta)h(w, A, X; \eta),\end{aligned}$$

in terms of θ and η. This modeling assumption is therefore appropriate only if the outcome mean admits the representation given above.

REMARK 7. As we show below, directly modeling the outcome confounding bridge function obviates the need to solve complicated integral equations which are well known to be ill-posed [1, 31] and therefore to admit unstable solutions. The above modeling strategy can be viewed as a form of regularization of the problem so as to resolve ill-posedness. However, a misspecified low dimensional model for say h can potentially lead to biased inference. In fact, inspired by the current paper, Cui et al. [14], Ying et al. [57] recently examined numerically the potential bias from model misspecification of h. As a potential remedy, they proposed a doubly robust estimator [14, 57] which provides extra robustness against possible model misspecification.

Although not pursued here, a variety of semiparametric (e.g., partially linear model, single index model) or nonparametric (e.g., generalized additive, reproducing kernels, neural networks) [20, 30] may be used to model h more flexibly, thus further alleviating concerns about specification bias.

Let $\widehat{\theta}$ denote the maximum likelihood estimator of θ, and define $\widehat{f}_{W|Z,A}(W) = f(W|Z, A, X; \widehat{\theta})$ implied by $f(L, A; \widehat{\theta})$. One may then estimate η based on Theorem 4.1, by fitting via least-squares, the regression model: $\mu(A, \eta) = E(Y|A, X, Z) = E(H(\eta)|A, X, Z)$ given by

$$\widehat{\mu}(A; \eta) = \sum_{w} \widehat{f}_{W|Z,A}(w) h(w, A, X; \eta). \tag{31}$$

For continuous Y, this may be accomplished by least-squares minimization [26],

$$\widehat{\eta} = \arg\min_{\eta} E_n\big\{\big(Y - \widehat{\mu}(A; \eta)\big)^2\big\},$$

where E_n stands for sample average. Then, assuming all models are correctly specified, one can show that $\widehat{\beta}(a)$ is a consistent and asymptotically normal estimator of $\beta(a)$, where

$$\widehat{\beta}(a) = E_n\big\{h(W, a, X; \widehat{\eta})\big\}.$$

For inference under the proposed parametric approach, we recommend using the nonparametric bootstrap to obtain standard errors and confidence intervals. However, we note that the bootstrap may not readily be justified when combined with data-adaptive and/or nonparametric estimators of nuisance parameters mentioned in above Remark 6. In such more flexible estimation settings, rate doubly robust estimators that accommodate slower than root-n estimation rates of nuisance parameters have been proposed with corresponding asymptotic theory justifying inference [14, 20]. We note that evaluating (31) might require evaluating either a sum, an integral or both with respect to a high dimensional variable w; in many cases, the sum/integral may not admit a closed form expression or may be computationally prohibitive to evaluate, in which case Monte Carlo approximation of $\widehat{\mu}(A, \eta)$ may provide a practical solution. In case of binary Y, use of a link function (say logit or probit link function) may be necessary in specifying a model for $h(w, A, X; \eta)$, in order to ensure that $\mu(A; \eta, \theta) = \Pr(Y = 1|A, Z, X; \eta, \theta)$ lies in the unit interval (0,1). Estimation in the binary case can then proceed by standard maximum likelihood estimation thus maximizing the log likelihood function

$$\widehat{\eta} = \arg\max_{\eta} E_n\big\{Y \log \widehat{\mu}(A; \eta) + (1 - Y) \log\big(1 - \widehat{\mu}(A; \eta)\big)\big\}.$$

In the Appendix, we describe special cases where $\widehat{\mu}(A, \eta)$ admits a closed form expression. Here, we discuss the important special case of proximal g-computation under a linear specification for $h(W, A, X; \eta)$, say

$$h(W, A, X; \eta) = \beta_a A + \eta_w' W + \eta_x' X,$$

so that $E(Y_a) = \beta(a) = \beta_0 + \beta_a a$ where $\beta_0 = \eta_w' E\{W\} + \eta_x' E\{X\}$ where $\eta_x' X$ includes an intercept term. Suppose further that one specifies a (multivariate) linear regression model

$$W = \big(1, Z', A, X'\big)\Theta + \varepsilon_W. \tag{32}$$

Then, one can estimate the average causal effect $\beta_a = E(Y_{a+1} - Y_a)$ with the regression coefficient $\widehat{\beta}_a$ obtained by fitting the standard linear regression model

$$Y = \beta_a A + \eta_w' \widehat{W} + \eta_x' X + \varepsilon_Y, \tag{33}$$

by least-squares, where $\widehat{W} = (1, Z', A, X')\widehat{\Theta}$ is the element-wise least-squares regression of W on $(1, Z', A, X')$ [37]. This procedure is motivated by (2). We refer to this procedure as proximal two-stage least squares (P2SLS) given its close relationship to 2SLS estimation in instrumental variable setting [56]. This connection in fact has implications for practice as it indicates that the estimator can be implemented with any off-the-shelf instrumental variable software, for example, "ivreg" [18] in R, which can perform 2SLS for multivariate exposure variable upon taking W as the endogenous (multivariate) variable, Z playing the role of IV, with A taken as a covariate. Such software can be used to obtain $\widehat{\beta}_a$ and corresponding confidence intervals, accounting for the uncertainty in the first stage estimator $\widehat{W}$. The model given in equation (33) has an interesting interpretation as it emulates a standard regression adjustment of confounding by X, and further adjusts for $\widehat{W}$ as proxy for the unmeasured factor U therefore deconfounding the standard regression approach. As mentioned in the introduction, we can therefore refer to $\widehat{W}$ as *proximal control variable* [49, 56].

REMARK 8. One may note from the description of P2SLS that in the event that $\dim(Z) < \dim(W)$, η_w in (33) may not be uniquely identified; nonetheless, it is straightforward to verify that all least squares solutions for $\widehat{\eta}_w$ yield a consistent estimator $\widehat{\beta}_a$. Either way, a test for the null hypothesis that there is no unmeasured confounding bias can be operationalized by a standard statistical test of the null hypothesis that all components of η_w are identically zero, which is readily available even when $\dim(Z) < \dim(W)$.

REMARK 9. Note that the linear specification of the bridge function above is implied and therefore compatible with the linear structural equation model (1) which rules out any interaction between A and U; however, the reverse does not hold, the linear bridge function does not a priori rule out such interaction.

REMARK 10. Although the above exposition of P2SLS focuses on a linear main effects only specification of the brigde function, the P2SLS approach can easily accommodate effect heterogeneity and nonlinearity by incorporating corresponding interactions and nonlinear transformations of variables to the bridge function model. For instance, one may include in the bridge function terms such $A \times X$ and $A \times W$ interactions as well as nonlinear transformations, say W^2 in the case of scalar W. The first stage regression would then be modified by fitting via OLS a linear regression for each term involving W on (A, Z, X); that is, a separate regression for W, $A * W$ and W^2 respectively. The second stage would then estimate the bridge function upon replacing interactions and nonlinear terms with corresponding fitted values from the first stage. For model selection, one may then assess the extent to which the estimated causal effect changes with increased model complexity of the bridge function. An analogous approach can be develop for binary outcome incorporating an appropriate link function.

Next, we consider the longitudinal setting where we observe an i.i.d sample of size n on $(\overline{A}, \overline{L} = (\overline{X}, \overline{Z}, \overline{W}))$. Then, parametric proximal g-computation relies on specifying parametric models for the outcome bridge functions

$$h_1(\overline{W}, \overline{A}, \overline{X}) = h_1\big(\overline{W}, \overline{A}, \overline{X}(j); \eta_1\big),$$

$$h_0\big(W(0), \overline{A}, X(0)\big) = h_0\big(W(0), \overline{A}, X(0); \eta_0\big),$$

with unknown parameter η_j; and for the joint law

$$f(\overline{L}, \overline{A}) = f(\overline{L}, \overline{A}; \theta),$$

with unknown parameter θ. Let $\widehat{\theta}$ denote the maximum likelihood estimator of θ, and define $\widehat{f}_1(\overline{W}) = f(\overline{W}|\overline{Z}, \overline{A}, \overline{X}; \widehat{\theta})$ and $\widehat{f}_0(W(0)) = f(w(0)|A(0), X(0), Z(0); \widehat{\theta})$ both deduced from $f(\overline{L}, \overline{A}; \widehat{\theta})$. Then we propose to estimate η_j based on Theorem 4.1, by recursively fitting regression models of Y on $\mu_1(\overline{A}; \theta, \eta_1) = E(H_1(\eta_1)|\overline{A}, \overline{X}, \overline{Z}; \theta)$, and of $H_1(\eta_1)$ on $\mu_0(\overline{A}; \eta_0, \theta) = E(H_0(a_1, \eta_0)|A(0), X(0), Z(0); \theta)$ given by

$$\widehat{\mu}_1(\eta_1) = \sum_{\overline{w}} \widehat{f}_1(\overline{w}) h_1(\overline{w}, \overline{A}, \overline{X}; \eta_1),$$

$$\widehat{\mu}_0\big(A(1); \eta_0\big) = \sum_{w(0)} \widehat{f}_0\big(w(0)\big) h_0\big(w(0), \overline{A}, X(0); \eta_0\big).$$

Each of these regressions can readily be performed via ordinary least-squares. Then, assuming all models are correctly specified and Theorem 4.1 holds, one can show that $\widehat{\beta}(\overline{a})$ is a consistent estimator of $\beta(\overline{a})$, and is approximately normally distributed, where

$$\widehat{\beta}(\overline{a}) = E_n\big\{h_0\big(W(0), \overline{a}, X(0); \widehat{\eta}_0\big)\big\}.$$

In order to estimate standard errors for $\widehat{\beta}(\overline{a})$ and confidence intervals for $\beta(\overline{a})$, we recommend using the nonparametric bootstrap [16]. We refer to the above estimation procedure as parametric proximal g-computation, the proximal analog to parametric g-computation algorithm of Robins [26, 41].

The algorithm simplifies tremendously in case of additive confounding bridge functions, say

$$h_1\big(\overline{W}, \overline{A}, \overline{X}(j); \eta_1\big) = \big(1, \mathrm{c}_a(\overline{A}), \mathrm{c}_w(\overline{W})', \mathrm{c}_w(\overline{X})', X'(0)\big)\eta_1, \tag{34}$$

$$h_0\big(W(0), \overline{A}, X(0); \eta_0\big) = \big(1, \mathrm{c}_a(\overline{A}), W'(0), X'(0)\big)\eta_0, \tag{35}$$

where for time-varying variable $B(j)$, $\mathrm{c}_b(\overline{B})$ denotes a user specified function of $\overline{B}$, for instance, we might take $\mathrm{c}_b(\overline{B}) = \mathrm{cum}(\overline{B}) = B(0) + B(1)$, where in case of vector B, the sum applies entry-wise such that $\mathrm{c}_b(\overline{B})$ is a vector of the same dimension as $B(j)$. Then proximal g-computation can be implemented by the following recursive least-squares algorithm (Algorithm 1).

It is important to note that although additive, the specific form of models used in Steps 1–4 is quite flexible and can accommodate both nonlinearities (e.g., using either polynomial or splines to model covariates) as well as interactions with treatment or among covariates. More flexible models can be somewhat more involved as each nonlinear specification of $\overline{W}$ entries requires a corresponding regression in Step 1. We also note that the particular manner in which a treatment or covariate history enters a given model is entirely to the discretion of the analyst. For instance, natural options for $\mathrm{c}_a(\overline{A})$ include $(A(0), A(1), A(0) \times A(1))$, $\mathrm{cum}(\overline{A})$ or simply $A(1)$ as viable alternatives depending on their respective goodness-of-fit.

Interestingly, under linearity of H_0 and H_1 with respect to $\overline{W}$, the proximal recursive least squares algorithm yields an estimator of $\beta(\overline{a})$, that remains consistent even if linear models (36) and (37) for $\overline{W}$ and $W(0)$ are misspecified. Likewise, in the point treatment setting, $\widehat{\beta}(a)$

Algorithm 1: Proximal recursive least squares algorithm

Step 1: fit the multivariate linear regression

$$\mathrm{c}_w(\overline{W}) = \big(1, \mathrm{c}_z(\overline{Z})', \mathrm{c}_a(\overline{A})', \mathrm{c}_x(\overline{X})', X'(0)\big) \times \Theta_1 + \boldsymbol{\varepsilon}_W \tag{36}$$

by applying least-squares separately to each entry of vector $\mathrm{c}(\overline{W})$, and let

$$\widehat{\mathrm{c}}_w = \big(1, \mathrm{c}_z(\overline{Z})', \mathrm{c}_a(\overline{A})', \mathrm{c}_x(\overline{X})', X'(0)\big)\widehat{\Theta}_1$$

denote its fitted values;

Step 2: fit the linear regression

$$Y = \big(1, \mathrm{c}_a(\overline{A})', \widehat{\mathrm{c}}_w', \mathrm{c}_x(\overline{X})', X'(0)\big)\eta_1 + \boldsymbol{\varepsilon}_Y$$

by least-squares where we note that $\widehat{\mathrm{c}}_w(\overline{W})$ has been substituted in for $\mathrm{c}(\overline{W})$, and let

$$\widehat{H}_1(\overline{A}) = \big(1, \mathrm{c}_a(\overline{A})', \mathrm{c}_w(\overline{W})', \mathrm{c}_x(\overline{X})', X'(0)\big)\widehat{\eta}_1;$$

Step 3: fit the multivariate linear regression

$$W(0) = \big(1, Z(0)', \mathrm{c}_a(\overline{A})', X(0)'\big)\Theta_0 + \boldsymbol{\varepsilon}_W \tag{37}$$

by applying least-squares separately to each entry of vector $W(0)$, and let

$$\widehat{W}(0) = \big(1, Z(0)', \mathrm{c}_a(\overline{A})', X(0)'\big)\widehat{\Theta}_0$$

denote its fitted values; fit the linear regression

$$\widehat{H}_1 = \big(1, \mathrm{c}_a(\overline{A})', \widehat{W}'(0), X'(0)\big)\eta_0 + \varepsilon_{h_1}$$

by least-squares, to obtain an estimate of H_0,

$$\widehat{H}_0(\overline{A}) = \big(1, \mathrm{c}_a(\overline{A})', W'(0), X'(0)\big)\widehat{\eta}_0;$$

Step 4: Evaluate

$$\widehat{\beta}(\overline{a}) = E_n\big\{\widehat{H}_0(\overline{a})\big\} = \big(1, \mathrm{c}_a(\overline{a})', E_n\big\{W'(0)\big\}, E_n\big\{X'(0)\big\}\big)\widehat{\eta}_0.$$

can be shown to remain consistent even if model (32) is not correctly specified provided that h_0 and h_1 are correctly specified. This property is similar to the robustness of the 2SLS estimator in instrumental variable methodology, which remains consistent provided that the second stage linear model is correctly specified, even if the first stage model is misspecified [56, 53]. The implication of this result is that OLS provides extra protection against model misspecification bias in modeling $\overline{W}$ as a linear model, including for binary or discrete components of W. This is an important property that does not generally hold for proximal g-computation algorithm which requires in addition to correct specification of h_0 and h_1, that one also specify $f(\overline{L}, \overline{A}; \theta)$ correctly. It is however possible to obtain an estimator of $h_0(\eta_0)$ and $h_1(\eta_1)$ using a recursive generalized methods of moments (RGMM) which does not require a model for $f(\overline{L}, \overline{A})$ and therefore is not susceptible to bias due to modeling the latter incorrectly. We refer the interested reader to Miao, Shi and Tchetgen Tchetgen [37] in point treatment case. A detailed treatment of this more robust estimation approach in longitudinal settings is described in Ying et al. [57]. It is worth noting that when $f(\overline{L}, \overline{A}; \theta)$ is correctly specified, one can generally expect proximal g-computation to be more efficient than proximal recursive least square and recursive generalized method of moments.

6. DATA APPLICATIONS

6.1 Point Treatment Application

We first illustrate proximal estimation of causal effects in a point treatment application to the Study to Understand Prognoses and Preferences for Outcomes and Risks of Treatments (SUPPORT) with the aim of evaluating the causal effect of right heart catheterization (RHC) during the initial care of critically ill patients in the intensive care unit (ICU) on survival time up to 30 days [13]. RHC was performed in 2184 patients within the initial 24 hours of ICU stay, while 3551 patients were managed without RHC. The SUPPORT study collected rich patient information encoded in 71 covariates, including demographics (such as age, sex, race, education, income, and insurance status), estimated probability of survival, comorbidity, vital signs, physiological status, and functional status. The outcome of interest is the number of days between admission and death or censoring at 30 days (Y). Ten variables measuring the patient's overall physiological status were measured from a blood test during the initial 24 hours in the ICU: `serum sodium`, `serum potassium`, `serum creatinine`, `bilirubin`, `albumin`, `PaO2/(0.01 * FiO2) ratio`, `PaCO2`, `serum PH (arterial)`, `white blood cell count`, and `hematocrit`. These variables may be subject to substantial measurement error and as single snapshot of underlying physiological state over time may be viewed as potential confounding proxies. Among the ten physiological status measures, four (`pafi1`, `paco21`, `ph1`, `hema1`) are strongly correlated with both the treatment and the outcome; thus we construct proxies Z and W from this reduced set of variables and collect all 67 remaining variables as covariates (X). We hypothesize that our four candidate proxies are equally likely to be valid treatment-inducing proxy or outcome-inducing proxy, we consider a practical strategy for allocating the four candidate proxies to bucket types b and c. The approach first ranks proxies according to their strength of association in treatment (based on logistic regression of A on $L = (X, Z, W)$) and outcome models (based on linear regression of Y given A

and $L = (X, Z, W)$) respectively; next, we select proxies in decreasing order of strength of association, first selecting the proxy with the strongest association with the outcome as outcome-inducing proxy and likewise for the treatment. In case of a tie, that is if these are the same variable, one may either decide to prioritize one of the two buckets or alternatively to randomize allocation to a proxy bucket type; upon allocating a given variable, say to the outcome-inducing proxy bucket, one subsequently removes the variable from the list of remaining treatment-inducing candidate proxies, and vice-versa. The algorithm stops when all proxies have been allocated. The algorithm produced the allocation $Z = (\texttt{pafi1}, \texttt{paco21})$ and $W = (\texttt{ph1}, \texttt{hema1})$. It is important to note that the above procedure primarily aims to identify relevant proxies, for which there is empirical evidence of an association with both outcome and treatment variables. However, relevant proxies cannot empirically be shown to satisfy the key exclusion restrictions required of valid proxies, for example, one cannot empirically rule out a causal effect of empirically relevant treatment proxies on either W or Y. Ultimately, validity of relevant proxies must be judged from subject-matter knowledge. In the current setting, physiological biomarkers are known to be measurement error prone a formal basis for judging their validity as proxies. For estimation of the causal effect of interest, we assume a linear outcome confounding bridge function,

$$h(a, x, w) = \eta_0 + \eta_a a + \eta_x' x + \eta_w' w,$$

in which case, the coefficient $\eta_a = \beta_a = E\{h(1, X, W) - h(0, X, W)\}$ encodes the causal effect of interest. After allocating the proxies, we apply two stage least squares to estimate the confounding bridge function, which can be implemented via routine R software such as `ivreg`. The following is a standard call of `ivreg` in R,

$$\text{ivreg}(\text{Y} \sim \text{A} + \text{X} + \text{W}|\text{A} + \text{X} + \text{Z}, \text{data} = \text{rhc}).$$

Ordinary least squares results in a negative and statistically significant causal effect estimate $\widehat{\beta}_a(\text{OLS}) = -1.25$ with standard error $= 0.28$. Outcome-inducing proxy `ph1` is associated with confounding bridge parameter ($\widehat{\eta}_w = -16.92$ standard error $= 8.8$), indicating moderate empirical evidence that unmeasured confounding might be biasing $\widehat{\beta}_a(\text{OLS})$. The causal effect estimate obtained by P2SLS is substantially larger than standard OLS point estimate $\widehat{\beta}_a(\text{Proximal}) = -1.80$ with corresponding standard error $= 0.43$. These results suggest that RHC may have an even more harmful effect on 30 day-survival among critically ill patients admitted into an ICU than previously documented. Results from this analysis are summarized in tables provided in the Supplemental Material.

6.2 Time-Varying Treatment Application

We reanalyze data from an article published by [12] on the potential protective effects of the anti-rheumatic therapy Methotrexate (MTX) among patients with rheumatoid arthritis. While Choi et al. [12] focused on survival as an endpoint and used a Cox marginal structural models to quantify joint treatment effects under SRA, here we consider the joint causal effects of MTX on average of reported number of tender joints, an important measure of disease progression, without appealing to SRA. Our analysis includes individuals who were older than age 18 years and who attended the Wichita Arthritis Center at least twice between Jan 1, 1981 (when weekly low-dose methotrexate therapy and health assessment questionnaire scores became available) and Dec 31, 1999; had rheumatoid arthritis fulfilling the 1958–1987 American College of Rheumatology (formerly the American Rheumatism Association) criteria for rheumatoid arthritis; and had not received methotrexate before their first visit to the center, who survived more than 12 months.

Methotrexate use and dose was recorded in the computer database at each clinic visit. We classified methotrexate exposure status as ever-treated or never-treated, that is, once a patient starts methotrexate therapy, he or she was considered on therapy for the rest of the follow-up. This approach provides a conservative estimate of methotrexate efficacy just as the intent-to-treat analysis does in a randomized clinical trial.

A thousand and ten patients with rheumatoid arthritis met our inclusion criteria, 183 of them were treated with methotrexate at month 6 of follow-up. We have recorded baseline covariates including age, sex, past smoking status, education level, rheumatoid arthritis duration, calendar year, and rheumatoid factor positive. Time-varying covariates included current smoking status, health assessment questionnaire, number of tender joints, patient's global assessment, erythrocyte sedimentation rate, number of disease-modifying antirheumatic drugs taken, and prednisone use. Our objective is therefore to evaluate the joint effects of MTX use at baseline and month six on average of tender joints at month 12 of follow-up. In addition to proximal causal inference, for comparison, similar to Choi et al. [12], we also evaluated the causal effect of interest under a marginal structural linear model $E(Y_{\overline{a}}) = \beta_0 + \beta_a \text{cum}(\overline{a}) = \beta_0 + \beta_a\{a(0) + a(1)\}$ where $a(0)$ and $a(1)$ are MTX use at baseline and at month 6 respectively, estimated via standard inverse probability weighted least squares assuming SRA given both all baseline and time-varying covariates.

We then implemented proximal recursive least squares algorithm under linear outcome confounding bridge specification (34) and (35), with $X =$ (age, education, sex, smoking, rheumatoid arthritis duration, calendar year). Since number of tender joints at one year of follow-up

is the primary outcome, tender joints count (`jc`) at baseline and at follow-up month 6 are both natural candidates as outcome-inducing proxies. Other candidate proxies included health assessment questionnaire (`haqc`), patient's global assessment of disease status (`gsc`) and erythrocyte sedimentation rate (`esrc`), number of disease-modifying antirheumatic drugs (`dmrd`), rheumatoid factor positive (`rapos`) and prednisone use (`onprd2`). We further reduced the set of candidate proxies to candidate variables associated with both treatment and outcome variables. Finally, we applied the allocation algorithm described in the prior section resulting in $Z(j) = \texttt{haqc}(j)$ and $W(j) = \texttt{jc}(j)$.

IPW least squares [41] suggests a protective effect of MTX with $\widehat{\beta}_{\overline{a}} = -0.23$ $(-0.43, -0.02)$, although validity of this finding is contingent on SRA. Proximal recursive least-squares yields results suggests a stronger protective effect $\widehat{\beta}_{\overline{a}} = -0.37$ $(-0.67, -0.13)$, with strong evidence of confounding bias $(\widehat{\eta}_{w,0}, \widehat{\eta}_{w,1}) = (0.785, 0.524)$ with corresponding 95% confidence intervals $(0.50, 1.1)$ and $(0.33, 0.71)$, respectively. These results reinforce understanding of potential protective effects of MTX on disease progression. Results from this analysis are summarized in tables provided in the Supplemental Material. We also report in the Supplemental Material a sensitivity analysis in which we vary which variables are used as W and Z, respectively, to illustrate the impact of classifying proxies.

7. DISCUSSION

We have described a new framework for the analysis of observational data subject to potential confounding bias. The approach acknowledges that in practice, measured covariates generally fail in observational settings to capture all potential confounding mechanisms and at most may be seen as proxy measurements of underlying confounding factors. Our proximal causal inference framework provides a formal potential outcome framework under which one can articulate conditions to identify causal effects from proxies. We have described proximal g-formula and proximal g-computation algorithm for estimation in point treatment and time-varying treatment settings. The proximal approach is closely related to negative control methods recently proposed for detection of hidden confounding bias and potentially for identification of causal effects of a point treatment intervention [33, 36, 37, 45]. We refer the reader to Shi, Miao and Tchetgen Tchetgen [46] for a recent review of negative control literature.

While similar to standard g-computation, our proximal g-computation algorithm (as well as proximal two-stage least squares and recursive least squares) rely on correct specification of outcome confounding bridge functions, we have also recently developed alternative methods [14, 20, 57] which similar to inverse-probability weighting, rely on a model for a so-called treatment confounding bridge function such that it is possible to construct two separate estimators of the average treatment effect each depending on a different model; either outcome or treatment confounding bridge function. Interestingly, we have also developed doubly robust estimators [14, 20, 57] that, similar to standard doubly robust estimators developed by Robins and colleagues [44], remain unbiased in large samples provided at least one confounding bridge function model is correct, but not necessarily both. As pointed out by a reviewer, the g-null paradox is known to potentially occur in longitudinal settings such as considered in this paper under a standard parametrization of the observed data likelihood. Robins has previously proposed several strategies to avoid the g-null paradox under sequential exchangeability (i.e., sequential unconfoundedness), typically involving a nonstandard parametrization of the observed data likelihood, with specific parameters encoding various causal effects in view. The proximal g-computation methods proposed in this paper also provide a non-standard parametrization of the observed data likelihood which likewise avoids the g-null paradox even when sequential exchangeability fails, but valid proxies are available. Specifically, the model parameters for specified outcome confounding bridge functions directly encode g-null hypotheses that may be of scientific interest thus effectively resolving concerns about the g-null paradox in complex longitudinal settings with unmeasured confounding.

ACKNOWLEDGMENTS

The authors would like to thank the anonymous referees, an Associate Editor and the Editor for their constructive comments that improved the quality of this paper.

The authors thank Dr. Stephen R. Cole for helpful comments.

SUPPLEMENTARY MATERIAL

Supplementary Material to "An Introduction to Proximal Causal Inference" (DOI: 10.1214/23-STS911SUPP; .pdf). Supplementary material includes proofs, tables, and additional results.

REFERENCES

[1] AI, C. and CHEN, X. (2003). Efficient estimation of models with conditional moment restrictions containing unknown functions. *Econometrica* **71** 1795–1843. MR2015420 https://doi.org/10.1111/1468-0262.00470

[2] AN, Y. and HU, Y. (2012). Well-posedness of measurement error models for self-reported data. *J. Econometrics* **168** 259–269. MR2923767 https://doi.org/10.1016/j.jeconom.2012.01.036

[3] ANDREWS, D. W. K. (2017). Examples of L^2-complete and boundedly-complete distributions. *J. Econometrics* **199** 213–220. MR3681027 https://doi.org/10.1016/j.jeconom.2017.05.011

[4] BLUNDELL, R., CHEN, X. and KRISTENSEN, D. (2007). Semi-nonparametric IV estimation of shape-invariant Engel curves. *Econometrica* **75** 1613–1669. MR2351452 https://doi.org/10.1111/j.1468-0262.2007.00808.x

[5] CANAY, I. A., SANTOS, A. and SHAIKH, A. M. (2013). On the testability of identification in some nonparametric models with endogeneity. *Econometrica* **81** 2535–2559. MR3138554 https://doi.org/10.3982/ECTA10851

[6] CARROLL, R. J., CHEN, X. and HU, Y. (2010). Identification and estimation of nonlinear models using two samples with nonclassical measurement errors. *J. Nonparametr. Stat.* **22** 379–399. MR2662599 https://doi.org/10.1080/10485250902874688

[7] CASELLA, G. and BERGER, R. L. (2001). *Statistical Inference*. *The Wadsworth & Brooks/Cole Statistics/Probability Series*. Wadsworth & Brooks/Cole Advanced Books & Software, Pacific Grove, CA. MR1051420

[8] CHEN, X., CHERNOZHUKOV, V., LEE, S. and NEWEY, W. K. (2014). Local identification of nonparametric and semiparametric models. *Econometrica* **82** 785–809. MR3191719 https://doi.org/10.3982/ECTA9988

[9] CHEN, X. and CHRISTENSEN, T. M. (2018). Optimal sup-norm rates and uniform inference on nonlinear functionals of nonparametric IV regression. *Quant. Econ.* **9** 39–84. MR3789729 https://doi.org/10.3982/QE722

[10] CHEN, X. and HU, Y. (2006). Identification and inference of nonlinear models using two samples with arbitrary measurement errors. Technical report, Cowles, Foundation for Research in Economics, Yale Univ.

[11] CHERNOZHUKOV, V. and HANSEN, C. (2005). An IV model of quantile treatment effects. *Econometrica* **73** 245–261. MR2115636 https://doi.org/10.1111/j.1468-0262.2005.00570.x

[12] CHOI, H. K., HERNÁN, M. A., SEEGER, J. D., ROBINS, J. M. and WOLFE, F. (2002). Methotrexate and mortality in patients with rheumatoid arthritis: A prospective study. *Lancet* **359** 1173–1177.

[13] CONNORS, A. F., SPEROFF, T., DAWSON, N. V., THOMAS, C., HARRELL, F. E., WAGNER, D., DESBIENS, N., GOLDMAN, L., WU, A. W. et al. (1996). The effectiveness of right heart catheterization in the initial care of critically III patients. *JAMA* **276** 889–897.

[14] CUI, Y., PU, H., SHI, X. and MIAO, W. (2023). Semiparametric proximal causal inference. *J. Amer. Statist. Assoc.* 1–12.

[15] DAROLLES, S., FAN, Y., FLORENS, J. P. and RENAULT, E. (2011). Nonparametric instrumental regression. *Econometrica* **79** 1541–1565. MR2883763 https://doi.org/10.3982/ECTA6539

[16] EFRON, B. and TIBSHIRANI, R. J. (1994). *An Introduction to the Bootstrap*. *Monographs on Statistics and Applied Probability* **57**. CRC Press, New York. MR1270903 https://doi.org/10.1007/978-1-4899-4541-9

[17] FOLSTEIN, M. F., FOLSTEIN, S. E. and MCHUGH, P. R. (1975). "Mini-mental state": A practical method for grading the cognitive state of patients for the clinician. *J. Psychiatr. Res.* **12** 189–198.

[18] FOX, J., KLEIBER, C. and ZEILEIS, A. (2021). ivreg: Instrumental-variables regression by '2SLS', '2SM', or '2SMM', with diagnostics. R package version 0.6-1.

[19] FREYBERGER, J. (2018). Non-parametric panel data models with interactive fixed effects. *Rev. Econ. Stud.* **85** 1824–1851. MR3818061 https://doi.org/10.1093/restud/rdx052

[20] GHASSAMI, A., YING, A., SHPITSER, I. and TCHETGEN TCHETGEN, E. J. (2021). Minimax kernel machine learning for a class of doubly robust functionals. arXiv preprint. Available at arXiv:2104.02929.

[21] GREENLAND, S. and ROBINS, J. M. (1986). Identifiability, exchangeability, and epidemiological confounding. *Int. J. Epidemiol.* **15** 413–419.

[22] HALL, P., HOROWITZ, J. L. et al. (2005). Nonparametric methods for inference in the presence of instrumental variables. *Ann. Statist.* **33** 2904–2929. MR2253107 https://doi.org/10.1214/009053605000000714

[23] HEATON, R. K., FRANKLIN, D. R., ELLIS, R. J., MCCUTCHAN, J. A., LETENDRE, S. L., LEBLANC, S., CORKRAN, S. H., DUARTE, N. A., CLIFFORD, D. B. et al. (2011). HIV-associated neurocognitive disorders before and during the era of combination antiretroviral therapy: Differences in rates, nature, and predictors. *J. Neurovirology* **17** 3–16.

[24] HERNÁN, M. A., BRUMBACK, B. A. and ROBINS, J. M. (2002). Estimating the causal effect of zidovudine on CD4 count with a marginal structural model for repeated measures. *Stat. Med.* **21** 1689–1709. https://doi.org/10.1002/sim.1144

[25] HERNÁN, M. A. and ROBINS, J. M. (2006). Instruments for causal inference: An epidemiologist's dream? *Epidemiology* **17** 360–372. https://doi.org/10.1097/01.ede.0000222409.00878.37

[26] HERNÁN, M. A. and ROBINS, J. M. (2020). *Causal Inference*: *What If*. CRC Press/CRC, Boca Raton, FL.

[27] HOROWITZ, J. L. (2011). Applied nonparametric instrumental variables estimation. *Econometrica* **79** 347–394. MR2809374 https://doi.org/10.3982/ECTA8662

[28] HU, Y. and SCHENNACH, S. M. (2008). Instrumental variable treatment of nonclassical measurement error models. *Econometrica* **76** 195–216. MR2374986 https://doi.org/10.1111/j.0012-9682.2008.00823.x

[29] HU, Y. and SHUM, M. (2012). Nonparametric identification of dynamic models with unobserved state variables. *J. Econometrics* **171** 32–44. MR2970334 https://doi.org/10.1016/j.jeconom.2012.05.023

[30] KALLUS, N., MAO, X. and UEHARA, M. (2021). Causal inference under unmeasured confounding with negative controls: A minimax learning approach. arXiv preprint. Available at arXiv:2103.14029.

[31] KRESS, R., MAZ'YA, V. and KOZLOV, V. (1989). *Linear Integral Equations*. *Applied Mathematical Sciences* **82**. Springer, Berlin. MR1007594 https://doi.org/10.1007/978-3-642-97146-4

[32] KUROKI, M. and PEARL, J. (2014). Measurement bias and effect restoration in causal inference. *Biometrika* **101** 423–437. MR3215357 https://doi.org/10.1093/biomet/ast066

[33] LIPSITCH, M., TCHETGEN TCHETGEN, E. J. and COHEN, T. (2010). Negative controls: A tool for detecting confounding and bias in observational studies. *Epidemiology* **21** 383.

[34] MIAO, W., GENG, Z. and TCHETGEN TCHETGEN, E. J. (2018). Identifying causal effects with proxy variables of an unmeasured confounder. *Biometrika* **105** 987–993. MR3877879 https://doi.org/10.1093/biomet/asy038

[35] MIAO, W., HU, W., OGBURN, E. L. and ZHOU, X.-H. (2023). Identifying effects of multiple treatments in the presence of unmeasured confounding. *J. Amer. Statist. Assoc.* **118** 1953–1967. MR4646619 https://doi.org/10.1080/01621459.2021.2023551

[36] MIAO, W. and TCHETGEN TCHETGEN, E. J. (2017). Invited commentary: Bias attenuation and identification of causal effects with multiple negative controls. *Amer. J. Epidemiol.* **185** 950–953.

[37] MIAO, W., SHI, X. and TCHETGEN TCHETGEN, E. J. (2018). A confounding bridge approach for double negative control inference on causal effects. arXiv preprint. Available at arXiv:1808.04945.

[38] NEWEY, W. K. and POWELL, J. L. (2003). Instrumental variable estimation of nonparametric models. *Econometrica* **71** 1565–1578. MR2000257 https://doi.org/10.1111/1468-0262.00459

[39] OGBURN, E. L. and VANDERWEELE, T. J. (2013). Bias attenuation results for nondifferentially mismeasured ordinal and coarsened confounders. *Biometrika* **100** 241–248. MR3034338 https://doi.org/10.1093/biomet/ass054

[40] PEARL, J. (2009). *Causality*: *Models*, *Reasoning*, *and Inference*, 2nd ed. Cambridge Univ. Press, Cambridge. MR2548166 https://doi.org/10.1017/CBO9780511803161

[41] ROBINS, J. M. (1997). Causal inference from complex longitudinal data. In *Latent Variable Modeling and Applications to Causality* (*Los Angeles*, *CA*, 1994). *Lect*. *Notes Stat*. **120** 69–117. Springer, New York. MR1601279 https://doi.org/10.1007/978-1-4612-1842-5_4

[42] ROBINS, J. M. (1999). Marginal structural models. In *Proceedings of the Section on Bayesian Statistical Science* 1–10. Amer. Statist. Assoc., Alexandria. MR1766776 https://doi.org/10.1023/A:1005285815569

[43] ROSENBAUM, P. R. and RUBIN, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika* **70** 41–55. MR0742974 https://doi.org/10.1093/biomet/70.1.41

[44] SCHARFSTEIN, D. O., ROTNITZKY, A. and ROBINS, J. M. (1999). Adjusting for nonignorable drop-out using semiparametric nonresponse models. *J. Amer. Statist*. *Assoc*. **94** 1096–1120. MR1731478 https://doi.org/10.2307/2669923

[45] SHI, X., MIAO, W., NELSON, J. C. and TCHETGEN TCHETGEN, E. J. (2020). Multiply robust causal inference with double-negative control adjustment for categorical unmeasured confounding. *J. R. Stat. Soc. Ser. B. Stat. Methodol*. **82** 521–540. MR4084174 https://doi.org/10.1111/rssb.12361

[46] SHI, X., MIAO, W. and TCHETGEN TCHETGEN, E. J. (2020). A selective review of negative control methods in epidemiology. *Curr. Epidemiol. Rep*. 1–13.

[47] SHIU, J.-L. and HU, Y. (2013). Identification and estimation of nonlinear dynamic panel data models with unobserved covariates. *J. Econometrics* **175** 116–131. MR3061934 https://doi.org/10.1016/j.jeconom.2013.03.001

[48] SOFER, T., RICHARDSON, D. B., COLICINO, E., SCHWARTZ, J. and TCHETGEN TCHETGEN, E. J. (2016). On negative outcome control of unobserved confounding as a generalization of difference-in-differences. *Statist. Sci*. **31** 348–361. MR3552739 https://doi.org/10.1214/16-STS558

[49] STOKES, A. and MEHTA, N. K. (2013). Mortality and excess risk in US adults with pre-diabetes and diabetes: A comparison of two nationally representative cohorts, 1988–2006. *Popul. Health Metr.* **11** 1–7.

[50] TCHETGEN TCHETGEN, E. J., YING, A., CUI, Y., SHI, X. and MIAO, W. (2024). Supplement to "An introduction to proximal causal inference." https://doi.org/10.1214/23-STS911SUPP

[51] TOGLIA, J., FITZGERALD, K. A., O'DELL, M. W., MASTROGIOVANNI, A. R. and LIN, C. D. (2011). The mini-mental state examination and Montreal cognitive assessment in persons with mild subacute stroke: Relationship to functional outcome. *Arch. Phys. Med. Rehabil*. **92** 792–798.

[52] TOMBAUGH, T. N. and MCINTYRE, N. J. (1992). The mini-mental state examination: A comprehensive review. *J. Amer. Geriatr. Soc*. **40** 922–935.

[53] VANSTEELANDT, S. and DIDELEZ, V. (2018). Improving the robustness and efficiency of covariate-adjusted linear instrumental variable estimators. *Scand. J. Stat*. **45** 941–961. MR3884895 https://doi.org/10.1111/sjos.12329

[54] WANG, L. and TCHETGEN TCHETGEN, E. (2018). Bounded, efficient and multiply robust estimation of average treatment effects using instrumental variables. *J. R. Stat. Soc. Ser. B. Stat. Methodol*. **80** 531–550. MR3798877 https://doi.org/10.1111/rssb.12262

[55] WOODS, S. P., MOORE, D. J., WEBER, E. and GRANT, I. (2009). Cognitive neuropsychology of HIV-associated neurocognitive disorders. *Neuropsychol. Rev.* **19** 152–168.

[56] WOOLDRIDGE, J. M. (2010). *Econometric Analysis of Cross Section and Panel Data*, 2nd ed. MIT Press, Cambridge, MA. MR2768559

[57] YING, A., MIAO, W., SHI, X. and TCHETGEN TCHETGEN, E. J. (2023). Proximal causal inference for complex longitudinal studies. *J. R. Stat. Soc. Ser. B. Stat. Methodol*. **85** 684–704.