Biometrika (2021), **108**, 4, *pp.* 881–894
Printed in Great Britain
doi: 10.1093/biomet/asab003
Advance Access publication 10 May 2021

Estimation of local treatment effects under the binary instrumental variable model

BY LINBO WANG

Department of Statistical Sciences, University of Toronto, Toronto, Ontario M5S 3G3, Canada
linbo.wang@utoronto.ca

YUEXIA ZHANG

Department of Computer and Mathematical Sciences, University of Toronto, Toronto, Ontario M1C 1A4, Canada
yuexia.zhang@utoronto.ca

THOMAS S. RICHARDSON

Department of Statistics, University of Washington, Box 354322, Seattle, Washington 98195, U.S.A.
thomasr@u.washington.edu

AND JAMES M. ROBINS

Department of Epidemiology, Harvard T. H. Chan School of Public Health, 677 Huntington Avenue, Boston, Massachusetts 02115, U.S.A.
robins@hsph.harvard.edu

SUMMARY

Instrumental variables are widely used to deal with unmeasured confounding in observational studies and imperfect randomized controlled trials. In these studies, researchers often target the so-called local average treatment effect as it is identifiable under mild conditions. In this paper we consider estimation of the local average treatment effect under the binary instrumental variable model. We discuss the challenges of causal estimation with a binary outcome and show that, surprisingly, it can be more difficult than in the case with a continuous outcome. We propose novel modelling and estimation procedures that improve upon existing proposals in terms of model congeniality, interpretability, robustness and efficiency. Our approach is illustrated via simulation studies and a real data analysis.

Some key words: Causal inference; Model compatibility; Semiparametric efficiency; Variation independence.

1. INTRODUCTION

Unmeasured confounding is a common threat to drawing valid causal inferences. It can occur in observational studies as well as imperfect randomized controlled trials where participants may not comply with the assigned treatment. Instrumental variable methods which seek to address this issue are widely used in economics, biostatistics and epidemiology to estimate causal effects when unmeasured confounders may be present. Intuitively, an instrumental variable is a pre-treatment

covariate that is associated with the outcome only through its effect on the treatment. In practice, often the condition above is reasonable only after controlling for a set of baseline covariates.

Traditionally, instrumental variable methods have aimed to estimate average treatment effects (Wright & Wright, 1928; Goldberger, 1972). Identification of the average treatment effects, however, relies on untestable homogeneity assumptions involving unmeasured confounders (e.g., Hernán & Robins, 2006; Wang & Tchetgen Tchetgen, 2018). An alternative approach, proposed by Imbens & Angrist (1994) and Angrist et al. (1996), is to estimate the so-called local average treatment effect, which can be nonparametrically identified under a certain monotonicity assumption. In the noncompliance setting, local average treatment effects may be of interest in practice, since if the local effect indicates that treatment is advantageous then this can be used as an argument for increasing the incentives for taking the treatment.

The problem of estimating local average treatment effects has been studied extensively for continuous outcomes (e.g., Abadie et al., 2002; Abadie, 2003; Tan, 2006; Okui et al., 2012; Ogburn et al., 2015). However, as explained in detail in § 2, direct application of these methods to binary outcomes is often inappropriate. Furthermore, with the exception of those developed by Abadie (2003) and Ogburn et al. (2015), most existing methods focus only on the additive local average treatment effect, but not the multiplicative local average treatment effect; the latter is commonly of interest in cases with binary outcomes, as it measures the causal effect on the relative risk scale. In related work, the Wald-type estimator of Didelez et al. (2010) can be shown to be approximately equal to the multiplicative local average treatment effect under a monotonicity assumption (Clarke & Windmeijer, 2012).

In this paper, we propose novel estimating procedures for both the additive and the multiplicative local average treatment effects with a binary outcome. The proposed procedures (i) ensure that the posited models are variation-independent, and hence congenial to each other (Meng, 1994); (ii) ensure that the resulting estimates lie in the natural nontrivial parameter space; (iii) directly parameterize the local average treatment effect curves to improve interpretability and reduce the risk of model misspecification (Ogburn et al., 2015); and (iv) allow for efficient and truly doubly robust estimation of the causal parameter of interest. To the best of our knowledge, for the additive local average treatment effect, our procedure is the first that achieves objective (iv); for the multiplicative local average treatment effect, our procedure is the first that achieves (i), (iii) or (iv); see also Remark 3.

2. FRAMEWORK, NOTATION AND EXISTING ESTIMATORS

Consider the problem of causal effect estimation with a binary exposure indicator D and a binary outcome Y. Suppose that the effect of D on Y is subject to confounding by observed variables X as well as unobserved variables U. Following the potential outcome framework, we assume $D(z)$, the potential exposure if the instrumental variable were to take value z, to be well-defined. Similarly, we assume $Y(z, d)$, the outcome that would have been observed if a unit were exposed to d and the instrument had taken value z, to be well-defined. We assume that we also observe a binary instrumental variable Z that satisfies the following assumptions (Angrist et al., 1996).

Assumption 1 (Exclusion restriction). For all z and z', $Y(z, d) = Y(z', d) \equiv Y(d)$ almost surely.

Assumption 2 (Independence). We have that $Z \perp\!\!\!\perp (Y(d), D(z)) \mid X$, $d = 0, 1$, $z = 0, 1$.

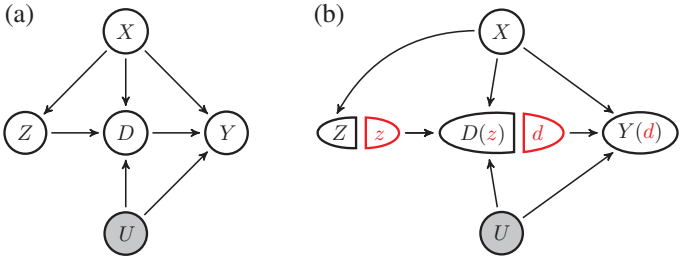

Fig. 1. Illustration of an instrumental variable model using a causal graph: (a) a causal directed acyclic graph (Pearl, 2009); (b) a single-world intervention graph (Richardson & Robins, 2013). Variables X, Z, D and Y are observed, and variable U is unobserved.

Table 1. *Principal strata t_D based on $\{D(1), D(0)\}$*

$D(1)$	$D(0)$	Principal stratum	Abbreviation
1	1	Always taker	AT
1	0	Complier	CO
0	1	Defier	DE
0	0	Never taker	NT

Assumption 3 (Instrumental variable relevance). We have that $\text{pr}\{D(1) = 1 \mid X\} \neq \text{pr}\{D(0) = 1 \mid X\}$ almost surely.

Assumption 4 (Positivity). There exists $\sigma > 0$ such that $\sigma < \text{pr}(Z = 1 \mid X) < 1 - \sigma$ almost surely.

Assumption 5 (Monotonicity). We have that $D(1) \geqslant D(0)$ almost surely.

Implicit in the notation $D(z)$ is that the instrument Z is causal so that Assumption 3 implies that Z has a nonzero causal effect on D. Figure 1 gives a simple illustration of the conditional instrumental variable model; see the Supplementary Material for another example.

Under the principal stratum framework (Frangakis & Rubin, 2002), the population can be divided into four strata based on values of $\{D(1), D(0)\}$, as shown in Table 1. We use t_D to denote principal strata defined by values of $\{D(1), D(0)\}$. We are interested in estimating the conditional treatment effects in the complier stratum on the additive and multiplicative scales, defined as

$$\begin{aligned} \text{LATE}(X) &= E\{Y(1) - Y(0) \mid D(1) > D(0), X\}, \\ \text{MLATE}(X) &= E\{Y(1) \mid D(1) > D(0), X\} / E\{Y(0) \mid D(1) > D(0), X\}. \end{aligned}$$

For $\text{MLATE}(X)$ to be well-defined, we also assume that $E\{Y(0) \mid D(1) > D(0), X\} \neq 0$ almost surely. By definition, the parameter spaces of $\text{LATE}(X)$ and $\text{MLATE}(X)$ are constrained: $\text{LATE}(X) \in [-1, 1]$ while $\text{MLATE}(X) \in [0, +\infty)$.

Abadie (2002, Lemma 2.1) showed that under Assumptions 1–5, the local average treatment effects are identifiable as

$$\text{LATE}(X) = \delta^{\text{L}}(X) \equiv \frac{E(Y \mid Z = 1, X) - E(Y \mid Z = 0, X)}{E(D \mid Z = 1, X) - E(D \mid Z = 0, X)}, \tag{1}$$

$$\text{MLATE}(X) = \delta^{\text{M}}(X) \equiv -\frac{E(YD \mid Z=1,X) - E(YD \mid Z=0,X)}{E\{Y(1-D) \mid Z=1,X\} - E\{Y(1-D) \mid Z=0,X\}}. \quad (2)$$

Given (1), it might be tempting to estimate $\delta^{\text{L}}(X)$, and hence $\text{LATE}(X)$ with a plug-in estimator by first estimating the four curves $E(Y \mid Z=z,X)$ and $E(D \mid Z=z,X)$ $(z=0,1)$ separately, as proposed by Frölich (2007). However, even though one may choose suitable models so that estimates for these four curves lie in the unit interval, there is no guarantee that the plug-in estimator will be between -1 and 1. The same problem arises when applying the approach of Tan (2006), which imposes parametric models on the conditional means $E(Y \mid D=d,Z=z,X)$ and $E(D \mid Z=z,X)$ $(d,z=0,1)$. Similar problems arise with plug-in estimators for $\delta^{\text{M}}(X)$.

To avoid these problems, Abadie (2003) proposed specifying a parametric model, such as the logistic model, for the so-called local average response function $E\{Y(d) \mid D(1) > D(0),X\}$. The model parameters are then estimated by a weighted estimating equation. The parameters of logistic models, however, do not directly encode the dependence of local average treatment effects on baseline covariates, so they do not offer direct insights into which baseline variables modify the local average treatment effects. Moreover, the validity of Abadie's (2003) approach hinges on correct specification of the instrumental density $\text{pr}(Z=1 \mid X)$. Instead, Okui et al. (2012) and Ogburn et al. (2015) proposed doubly robust estimators based on direct parameterization of the target functional $\delta^{\text{L}}(X)$. Given a correct model $\delta^{\text{L}}(X;\alpha)$, their estimators are consistent and asymptotically normal for the parameter of interest α if either the instrumental density model $\text{pr}(Z=1 \mid X;\gamma)$ or another nuisance model $E(Y - D \times \delta^{\text{L}}(X) \mid X;\beta)$ is correctly specified. However, the nuisance model $E(Y - D \times \delta^{\text{L}}(X) \mid X;\beta)$ is variation-dependent on the target model $\delta^{\text{L}}(X;\alpha)$. In this case, the double robustness properties of the estimators proposed by Okui et al. (2012) and Ogburn et al. (2015) are not practically meaningful, since with continuous covariates it is often not possible for $\delta^{\text{L}}(X;\alpha)$ and $E(Y - D \times \delta^{\text{L}}(X) \mid X;\beta)$ to be correct simultaneously. Similar considerations apply to the target functional $\delta^{\text{M}}(X)$.

In related work, Wang & Tchetgen Tchetgen (2018) studied the problem of estimating a closely related functional, $E_X\{\delta^{\text{L}}(X)\}$, which may be interpreted as the average treatment effect under a certain set of identification assumptions. As an intermediate step, Wang & Tchetgen Tchetgen (2018, § 4.1) proposed alternative nuisance models that are variation-independent of $\delta^{\text{L}}(X;\alpha)$, including a model for $\delta^{D}(X) \equiv E(D \mid Z=1,X) - E(D \mid Z=0,X)$. The key observation made by these authors is that as long as the models for $\delta^{\text{L}}(X)$ and $\delta^{D}(X)$ both lie in their respective parameter spaces, i.e., $[-1,1]$ for $\delta^{\text{L}}(X)$ and $[0,1]$ for $\delta^{D}(X)$, then

$$E(Y \mid Z=1,X) - E(Y \mid Z=0,X) = \delta^{\text{L}}(X) \times \delta^{D}(X)$$

also lies in its parameter space $[-1,1]$. Based on this, they derived a maximum likelihood estimator (Wang & Tchetgen Tchetgen, 2018, § 4.1) and a truly doubly robust estimator for $\delta^{\text{L}}(X)$ (Wang & Tchetgen Tchetgen, 2018, equation (14)). These approaches, however, cannot be adapted to estimate $\delta^{\text{M}}(X)$ or the multiplicative local average treatment effect. Furthermore, as we explain later in Remark 2, even for the additive local average treatment effect, in general their doubly robust estimator fails to achieve the semiparametric efficiency bound.

3. A NOVEL PARAMETERIZATION

In this section we describe a novel parameterization of the observed-data likelihood involving $\delta^{\text{L}}(X;\alpha)$ or $\delta^{\text{M}}(X;\alpha)$. Specifically, our goal is to find nuisance models such that (I) they are

variation-independent of each other; (II) they are variation-independent of $\delta^{\mathrm{L}}(X;\alpha)$ and $\delta^{\mathrm{M}}(X;\alpha)$; and (III) there exists a bijection between the observed-data likelihood on $\mathrm{pr}(D=d, Y=y \mid Z=z, X)$ and the combination of target and nuisance models. The remaining parts of the likelihood on $\mathrm{pr}(Z=z, X)$ do not show up in the identification formula (1) or (2). Thus they contain no information about the parameters of interest and need not be modelled.

Let $p_x(d,y \mid z) = \mathrm{pr}(D=d, Y=y \mid Z=z, X=x)$. For any x, the parameter space of the observed-data likelihood on $p_x(d,y \mid z)$ is a six-dimensional space in $[0,1]^8$ (Richardson et al., 2011),

$$\Delta = \Big\{ p_x(d,y \mid z) \geqslant 0 : \sum_{d,y} p_x(d,y \mid z) = 1, \\ p_x(1,y \mid 1) \geqslant p_x(1,y \mid 0),\ p_x(0,y \mid 1) \leqslant p_x(0,y \mid 0),\ y=0,1 \Big\}. \quad (3)$$

Parameterization of $p_x(d,y \mid z)$ is a difficult problem since, as shown by (3), the likelihood components $p_x(d,y \mid z), d, y, z = 0, 1$ are not variation-independent of each other.

To make progress, instead of modelling the observed likelihood components directly, we seek to model components of the potential outcome likelihood $\{D(1), D(0), Y(1), Y(0)\}$ conditional on X. Specifically, we will consider $p(\mathrm{AT};X) \equiv \mathrm{pr}(t_D = \mathrm{AT} \mid X)$, $p(\mathrm{NT};X)$ and $p(\mathrm{CO};X)$ as well as $p(Y(1) \mid \mathrm{AT};X) \equiv \mathrm{pr}(Y(1) = 1 \mid t_D = \mathrm{AT}, X)$, $p(Y(0) \mid \mathrm{NT};X)$, $p(Y(1) \mid \mathrm{CO};X)$ and $p(Y(0) \mid \mathrm{CO};X)$. The remaining parts of the potential outcome likelihood, such as $p(Y(1) \mid \mathrm{NT};X)$, are not modelled as they are not related to the observed-data likelihood, and hence contain no information about the parameters of interest. The modelled components, however, are still variation-dependent since

$$p(\mathrm{AT};X) + p(\mathrm{NT};X) + p(\mathrm{CO};X) = 1. \quad (4)$$

Moreover, they do not contain our target function $\delta^{\mathrm{L}}(X)$ or $\delta^{\mathrm{M}}(X)$, which we denote by $\theta(X)$.

Theorem 1 presents an alternative parameterization that achieves goals (I)–(III). To avoid the constraint (4), we follow Wang et al. (2017c) and reparameterize $p(\mathrm{AT};X)$, $p(\mathrm{NT};X)$ and $p(\mathrm{CO};X)$. To reparameterize $p(Y(1) \mid \mathrm{CO};X)$ and $p(Y(0) \mid \mathrm{CO};X)$ so that the new parameterization includes $\theta(X)$, we follow Richardson et al. (2017) to model an odds product function in the complier stratum. The proof of Theorem 1 is given in the Supplementary Material.

THEOREM 1. *Let $\mathcal{M}$ denote the six-dimensional models consisting of the target model $\theta(X;\alpha)$ and models on the following nuisance functions:*

$$\phi_1(X) \equiv \mathrm{pr}(t_D = \mathrm{CO} \mid X) = \mathrm{pr}(D=1 \mid Z=1, X) - \mathrm{pr}(D=1 \mid Z=0, X),$$
$$\phi_2(X) \equiv \mathrm{pr}(t_D = \mathrm{AT} \mid t_D \in \{\mathrm{AT}, \mathrm{NT}\}, X) = \frac{\mathrm{pr}(D=1 \mid Z=0, X)}{\mathrm{pr}(D=1 \mid Z=0, X) + \mathrm{pr}(D=0 \mid Z=1, X)},$$
$$\phi_3(X) \equiv \mathrm{pr}(Y=1 \mid t_D = \mathrm{NT}, X) = \mathrm{pr}(Y=1 \mid D=0, Z=1, X),$$
$$\phi_4(X) \equiv \mathrm{pr}(Y=1 \mid t_D = \mathrm{AT}, X) = \mathrm{pr}(Y=1 \mid D=1, Z=0, X),$$
$$\mathrm{OP}^{\mathrm{CO}}(X) \equiv \frac{E\{Y(1) \mid t_D = \mathrm{CO}, X\} E\{Y(0) \mid t_D = \mathrm{CO}, X\}}{[1 - E\{Y(1) \mid t_D = \mathrm{CO}, X\}][1 - E\{Y(0) \mid t_D = \mathrm{CO}, X\}]},$$

where $\mathrm{OP}^{\mathrm{CO}}$ denotes the odds product in the complier stratum.

Under Assumptions 1–5, for any realization of X, the map

$$\{\mathrm{pr}(D=d, Y=y \mid Z=z, X), d, y, z \in \{0,1\}\} \\ \to \{\theta(X), \phi_1(X), \phi_2(X), \phi_3(X), \phi_4(X), \mathrm{OP}^{\mathrm{CO}}(X)\} \tag{5}$$

is well-defined and is a smooth bijection from Δ to $\mathcal{D} \times [0,1]^4 \times [0,\infty)$, where $\mathcal{D} = [-1,1]$ if $\theta(X) = \delta^{\mathrm{L}}(X)$ and $\mathcal{D} = [0,\infty)$ if $\theta(X) = \delta^{\mathrm{M}}(X)$. Furthermore, the models in $\mathcal{M}$ are variation-independent of each other.

Suppose that models for $\theta(X), \phi_1(X), \ldots, \phi_4(X)$ and $\mathrm{OP}^{\mathrm{CO}}(X)$ are all specified up to a finite-dimensional parameter; then these parameters, and in particular the local average treatment effects, may be estimated directly via unconstrained maximum likelihood based on the diffeomorphism (5). Likelihood-based confidence intervals can then be obtained in a standard fashion.

Remark 1. Since the constituent models $\theta(X), \phi_1(X), \ldots, \phi_4(X)$ and $\mathrm{OP}^{\mathrm{CO}}(X)$ are variation-independent, the modeller is free to pick any function of X with the given range. For example, to mitigate model misspecification, one may assume flexible machine learning models on these functions of X. In this case, one can similarly fit these flexible models based on the implied models on the likelihood $\mathrm{pr}(D=d, Y=y \mid Z=z, X)$.

4. Doubly robust estimation

In this section, we use our parameterization in Theorem 1 to construct truly doubly robust estimators that are asymptotically linear for estimating the local average treatment effects if either the nuisance models $\phi_1(X;\beta_1), \ldots, \phi_4(X;\beta_4), \mathrm{OP}^{\mathrm{CO}}(X;\eta)$ or the instrumental density model $\mathrm{pr}(Z=1 \mid X;\gamma)$ is correct, given that the causal model $\theta(X;\alpha)$ is correctly specified. These estimators are said to be truly doubly robust because, as shown in Theorem 1, the nuisance models and causal model are variation-independent, and hence congenial to each other.

Let $\hat{\alpha}, \hat{\beta}_1, \ldots, \hat{\beta}_4, \hat{\eta}$ and $\hat{\gamma}$ be the maximum likelihood estimators of $\alpha, \beta_1, \ldots, \beta_4, \eta$ and γ, respectively. Also let

$$H(Y,D,X;\alpha) = \begin{cases} Y - D\theta(X;\alpha), & \theta(X) = \delta^{\mathrm{L}}(X), \\ Y\theta(X;\alpha)^{-D}, & \theta(X) = \delta^{\mathrm{M}}(X). \end{cases}$$

We have the following theorem.

Theorem 2. *Let $\hat{\alpha}_{\mathrm{dr}}$ solve the estimating equation*

$$\mathbb{P}_n\, \omega(X) \frac{2Z-1}{f(Z \mid X;\hat{\gamma})} \big[H(Y,D,X;\alpha) - \hat{E}\{H(Y,D,X;\alpha) \mid X\}\big] = 0, \tag{6}$$

where $\mathbb{P}_n$ denotes the empirical mean operator, $\omega(X)$ is an arbitrary measurable function of X,

$$f(Z \mid X;\hat{\gamma}) = \{\mathrm{pr}(Z=1 \mid X;\hat{\gamma})\}^Z \{1 - \mathrm{pr}(Z=1 \mid X;\hat{\gamma})\}^{1-Z},$$

$$\hat{E}\{H(Y,D,X;\alpha)\,|\,X\} = \begin{cases} \hat{f}_0\hat{\phi}_1 + (1-\hat{\phi}_1)(1-\hat{\phi}_2)\hat{\phi}_3 + (1-\hat{\phi}_1)\hat{\phi}_2\hat{\phi}_4 - \theta(1-\hat{\phi}_1)\hat{\phi}_2, & \theta(X) = \delta^{\mathrm{L}}(X), \\ \hat{f}_0\hat{\phi}_1 + (1-\hat{\phi}_1)\hat{\phi}_2\hat{\phi}_4\theta^{-1} + (1-\hat{\phi}_1)(1-\hat{\phi}_2)\hat{\phi}_3, & \theta(X) = \delta^{\mathrm{M}}(X), \end{cases}$$

with

$$\hat{f}_0 = \begin{cases} \dfrac{1}{2(\hat{\mathrm{OP}}-1)}\big[\hat{\mathrm{OP}}(2-\theta) + \theta - \{\theta^2(\hat{\mathrm{OP}}-1)^2 + 4\,\hat{\mathrm{OP}}\}^{1/2}\big], & \theta(X) = \delta^{\mathrm{L}}(X), \\ \dfrac{1}{2\theta(1-\hat{\mathrm{OP}})}\big[-(\theta+1)\hat{\mathrm{OP}} + \{\hat{\mathrm{OP}}^2(\theta-1)^2 + 4\theta\,\hat{\mathrm{OP}}\}^{1/2}\big], & \theta(X) = \delta^{\mathrm{M}}(X), \end{cases}$$

and

$$\theta = \theta(X;\alpha), \quad \hat{\phi}_i = \phi(X;\hat{\beta}_i)\ \ (i = 1,\ldots,4), \quad \hat{\mathrm{OP}} = \mathrm{OP}^{\mathrm{CO}}(X;\hat{\eta}).$$

Then under a correct model for $\theta(X;\alpha)$ and regularity conditions, $\hat{\alpha}_{\mathrm{dr}}$ is consistent and asymptotically normally distributed provided that at least one of the models for $E\{H(Y,D,X;\alpha)\,|\,X\}$ or $f(Z\,|\,X;\gamma)$ is correctly specified. The optimal choice of $\omega(X)$ that minimizes the asymptotic variance of $\hat{\alpha}_{\mathrm{dr}}$ is given in the Supplementary Material.

Theorem 2 is a special case of the doubly robust g-estimation theory developed by Ogburn et al. (2015, § § 3.2 and 3.3). For completeness, we provide the proof in the Supplementary Material. One can also easily verify that the arguments of the square roots in $\hat{f}_0$ are always nonnegative, provided that the estimates of $\theta(X)$ and $\mathrm{OP}^{\mathrm{CO}}(X)$ stay within their respective domains. Statistical inference may be based on standard M-estimation theory. Alternatively, in the simulations and real data analysis, we use the nonparametric bootstrap.

Remark 2. When $\theta(X) = \delta^{\mathrm{L}}(X)$, the approach of Wang & Tchetgen Tchetgen (2018) parameterizes the marginal distributions $\mathrm{pr}(Y = 1\,|\,Z = z,X)$ and $\mathrm{pr}(D = 1\,|\,Z = z,X)$, whereas our approach parameterizes the joint distribution $\mathrm{pr}(D = d, Y = y\,|\,Z = z,X)$. On the other hand, in addition to the marginal distributions, the optimal choice of $\omega(X)$ also depends on $\mathrm{pr}(DY = 1\,|\,Z = z,X)$. Hence it can be calculated based on our parameterization, but not that of Wang & Tchetgen Tchetgen (2018).

Remark 3. Prompted by the comment of a referee, we point out that if $\theta(X) = \delta^{\mathrm{L}}(X)$, then $E\{H(Y,D,X;\alpha)\,|\,X\} = E(Y\,|\,X) - \theta(X;\alpha)E(D\,|\,X)$. Hence a simple way to estimate α based on (6) is to first obtain estimates of $E(Y\,|\,X)$, $E(D\,|\,X)$ and $f(Z\,|\,X)$ and then plug these estimates into (6) to estimate α. Moreover, in the Supplementary Material we show that $E(Y\,|\,X), E(D\,|\,X), f(Z\,|\,X)$ and $\delta^{\mathrm{L}}(X)$ are variation-independent in the interior of their domains; a similar phenomenon was previously observed by Wang et al. (2017a, § 3.1). Similarly, if $\theta(X) = \delta^{\mathrm{M}}(X)$, then $E\{H(Y,D,X;\alpha)\,|\,X\} = \mathrm{pr}(D = 1\,|\,X)\,\mathrm{pr}(Y = 1\,|\,D = 1,X)\theta(X;\alpha)^{-1} + \mathrm{pr}(D = 0\,|\,X)\,\mathrm{pr}(Y = 1\,|\,D = 0,X)$. Hence a simple way to estimate α based on (6) is to first obtain estimates of $E(Y\,|\,X), E(Y\,|\,D,X)$ and $f(Z\,|\,X)$, and then plug these estimates into (6) to estimate α; $E(Y\,|\,X)$, $E(Y\,|\,D,X)$, $f(Z\,|\,X)$ and $\delta^{\mathrm{M}}(X)$ are also variation-independent in the interior of their domains. Consequently, since all the nuisance models can, logically, be correctly specified, these simple estimators are truly doubly robust assuming that the true parameter values are away from the boundary. However, like the estimator of Wang & Tchetgen Tchetgen (2018), they cannot be used to estimate the optimal $\omega(X)$. Furthermore, these simple parameterizations do not lead to likelihood-based inference.

5. SIMULATION STUDIES

In this section, we evaluate the finite-sample performance of various estimators discussed in this paper. We generate data from the following models:

$$\delta^{\mathrm{L}}(X) = \tanh(\alpha^{\mathrm{T}}X), \quad \delta^{\mathrm{M}}(X) = \exp(\alpha^{\mathrm{T}}X),$$
$$\phi_i(X) = \mathrm{expit}(\beta_i^{\mathrm{T}}X) \ \ (i = 1, \ldots, 4), \quad \mathrm{OP}^{\mathrm{CO}}(X) = \exp(\eta^{\mathrm{T}}X),$$
$$\mathrm{pr}(Z = 1 \mid X) = \mathrm{expit}(\gamma^{\mathrm{T}}X),$$

where the covariates X include an intercept and a random variable generated from $\mathrm{Un}(-1, 1)$, $\alpha = (0, -1)^{\mathrm{T}}$, $\beta_i = (-0.4, 0.8)^{\mathrm{T}}$ $(i = 1, \ldots, 4)$, $\eta = (-0.4, 1)^{\mathrm{T}}$ and $\gamma = (0.1, -1)^{\mathrm{T}}$. Under this setting, the strength of the instrumental variable, defined as $\Delta^D = E\{E(D \mid Z = 1, X) - E(D \mid Z = 0, X)\}$, is 0.406. The sample size is 1000.

We also consider scenarios in which the nuisance models are misspecified. In these scenarios, the analyst is given covariates $X^\dagger$ that include an intercept and an irrelevant covariate generated from an independent $\mathrm{Un}(-1, 1)$, as well as covariates X' including

$$(\underbrace{1, \ldots, 1}_{0.5n}, \underbrace{0 \ldots, 0}_{0.5n})^{\mathrm{T}}, \quad (\underbrace{0, \ldots, 0}_{0.1n}, \underbrace{1, \ldots, 1}_{0.9n})^{\mathrm{T}}.$$

Instead of formulating a model conditioning on X, the analyst fits the model $\mathrm{pr}(Z = 1 \mid X^\dagger; \gamma)$ and/or $\phi_i(X'; \beta_i)$ $(i = 1, \ldots, 4)$ and $\mathrm{OP}^{\mathrm{CO}}(X'; \eta)$. The analyst still uses the correct functional form in these models. The target model $\theta(X; \alpha)$ is always correctly specified. In the Supplementary Material we visualize the degree of model misspecification by plotting the data points generated under the true models and misspecified models from one randomly selected Monte Carlo run.

We assess the performance of the following estimators.

mle:	the proposed maximum likelihood estimator;
drw:	the proposed doubly robust estimator with the optimal weighting function;
dru:	the proposed doubly robust estimator with the identity weighting function;
reg.ogburn:	the outcome regression estimator of Ogburn et al. (2015, § 3.1);
drw.ogburn:	the doubly robust estimator of Ogburn et al. (2015, § 3.3) with the optimal weighting function;
dru.ogburn:	the doubly robust estimator of Ogburn et al. (2015, § 3.2) with the identity weighting function;
mle.wang:	the maximum likelihood estimator of Wang & Tchetgen Tchetgen (2018, § 4.1);
dru.wang:	the doubly robust estimator of Wang & Tchetgen Tchetgen (2018, § 4.4) with the identity weighting function;
dru.simple:	the doubly robust estimator described in Remark 3;
ls.abadie:	the least squares estimator of Abadie (2003, § 4.2.1);
mle.crude:	the maximum likelihood estimator of the crude association on the additive or multiplicative scale (Richardson et al., 2017, § 2).

For models other than those described above, we provide details of the model specifications in the Supplementary Material.

Throughout our simulations, we assume that the model of interest $\theta(X)$ is always correctly specified. We consider the following four scenarios for the nuisance models.

Table 2. *Bias* ($\times 100$) *and standard error* ($\times 100$, *in parentheses*) *of the estimated biases in the Monte Carlo study of various estimators in the selected scenarios; the true values of* α_0 *and* α_1 *are* 0 *and* -1, *respectively, and the sample size is* 1000

	$\theta(X) = \delta^{\mathrm{L}}(X)$				$\theta(X) = \delta^{\mathrm{M}}(X)$			
	α_0		α_1		α_0		α_1	
mle.bth	0.28	(0.35)	−3.5	(0.78)	−0.092	(0.71)	−3.0	(1.2)
mle.bad	−20	(0.42)	−15	(0.80)	−48	(1.2)	−18	(2.1)
drw.bth	0.55	(0.36)	−4.1	(0.82)	0.54	(0.77)	−5.6	(1.5)
drw.psc	0.060	(0.38)	−5.9	(1.0)	−0.38	(1.2)	−12	(2.7)
drw.opc	0.55	(0.36)	−3.9	(0.79)	0.49	(0.75)	−5.3	(1.4)
drw.bad	−10	(0.40)	−9.6	(1.1)	−28	(1.4)	25	(3.3)
dru.bth	1.3	(0.44)	−5.8	(1.0)	1.8	(0.84)	−8.1	(1.7)
reg.ogburn.bth	−5.7	(1.6)	−2.9	(3.1)	7.8	(2.0)	−1.1	(2.2)
reg.ogburn.bad	−9.0	(0.25)	100	(0.23)	140	(5.6)	93	(3.6)
drw.ogburn.bth	0.10	(0.46)	−4.2	(0.99)	3.2	(1.4)	−13	(2.5)
dru.ogburn.bth	1.3	(0.45)	−5.8	(1.1)	1.9	(0.85)	−8.2	(1.7)
dru.wang.bth	1.3	(0.45)	−5.8	(1.0)	—		—	
dru.simple.bth	1.3	(0.45)	−5.8	(1.0)	1.8	(0.84)	−8.0	(1.7)
dru.simple.psc	1.2	(0.44)	−6.2	(1.0)	1.9	(0.84)	−8.8	(1.7)
dru.simple.opc	4.5	(0.49)	−17	(1.2)	−0.15	(0.68)	11	(1.2)
dru.simple.bad	−16	(0.48)	−17	(1.3)	−34	(0.70)	18	(1.5)
ls.abadie.bth	−0.19	(0.37)	−4.1	(0.93)	0.42	(0.79)	−11	(1.6)
ls.abadie.bad	−23	(0.88)	22	(1.2)	−32	(1.9)	7.7	(3.6)
mle.crude	−2.8	(0.10)	60	(0.19)	0.36	(0.25)	51	(0.42)

bth: X is used in all nuisance models;

psc: X is used in the instrumental density model, and X' is used in other nuisance models;

opc: $X^{\dagger}$ is used in the instrumental density model, and X is used in other nuisance models;

bad: $X^{\dagger}$ is used in the instrumental density model, and X' is used in other nuisance models.

As the method of Abadie (2003) does not directly specify a model for $\theta(X)$, we consider the following two scenarios for this method.

bth: X is used in all models;

bad: $X^{\dagger}$ is used in the instrumental density model,
but X is used in the model for $E\{Y \mid X, D, D(1) > D(0)\}$.

The implied model for $\theta(X)$ remains correct in either of these two scenarios.

Table 2 presents selected results of the bias and Monte Carlo standard error for various estimators based on 1000 Monte Carlo runs. In the Supplementary Material we present the complete set of results and the bias as a percentage of the estimator's standard deviation. The estimator mle.crude has large bias, indicating that the effect of unmeasured confounding is nonnegligible. As expected, the proposed estimators have small bias relative to the standard error in all scenarios except for mle.bad and drw.bad. As expected, when all nuisance models are correctly specified, the proposed maximum likelihood estimator has standard error smaller than or comparable to that of the optimally weighted doubly robust estimator drw.bth. The performances of dru.ogburn.bth, dru.wang.bth and dru.simple.bth are all similar to that of dru.bth; all four of these methods are

Table 3. *Coverage probabilities* (×100) *of confidence intervals obtained from* 500 *bootstrap samples in selected scenarios; the true values of* α_0 *and* α_1 *are* 0 *and* −1*, respectively, and the sample size is* 1000

	$\theta(X) = \delta^{\mathrm{L}}(X)$		$\theta(X) = \delta^{\mathrm{M}}(X)$	
	α_0	α_1	α_0	α_1
mle.bth	95.6	95.8	95.4	96.4
mle.bad	65.4	91.0	46.3	94.6
drw.bth	94.7	95.2	96.9	95.9
drw.psc	95.4	95.5	97.6	97.4
drw.opc	95.0	95.6	96.1	96.1
drw.bad	87.0	95.3	91.8	96.8
dru.bth	94.5	94.6	96.3	96.9
reg.ogburn.bth	98.0	99.9	99.6	100.0
reg.ogburn.bad	75.6	0.1	99.9	86.1
drw.ogburn.bth	97.0	98.1	98.5	98.4
dru.ogburn.bth	94.5	95.0	96.2	97.2
dru.wang.bth	94.4	94.7	—	—
dru.simple.bth	94.3	94.8	96.1	96.5
dru.simple.psc	94.7	94.6	96.1	96.8
dru.simple.opc	93.6	92.6	96.2	94.9
dru.simple.bad	76.3	94.4	69.3	93.8
ls.abadie.bth	94.8	95.5	96.4	95.6
ls.abadie.bad	87.3	94.5	93.1	95.7
mle.crude	84.3	0.0	94.0	6.3

less efficient than drw.bth. This suggests that under our simulation settings, the optimal weighting function yields important efficiency gains. Although drw.ogburn.bth is constructed based on the same optimally weighted estimating equation as drw.bth, misspecification of variation-dependent models leads to a biased estimate of the optimal weight function. As a result, in some cases it is even less efficient than dru.ogburn.bth.

Table 3 reports selected coverage probabilities of 95% confidence intervals obtained from the quantile bootstrap based on 500 bootstrap samples; the complete set of results is presented in the Supplementary Material. The proposed estimators have coverage close to the nominal level except for mle.bad and drw.bad. Inference results produced by reg.ogburn.bth and drw.ogburn.bth tend to be overly conservative, possibly due to misspecification of variation-dependent models.

6. APPLICATION TO 401(K) DATA

We use the proposed procedures to evaluate the effect of the 401(k) retirement plan on savings. The 401(k) plan has become the most popular employer-sponsored retirement plan in the U.S.A. Economists have long been interested in whether 401(k) contributions represent additional savings or simply replace other retirement plans, such as Individual Retirement Accounts. To account for unobserved confounders such as the underlying preference for savings, Abadie (2003) chose to use 401(k) eligibility as an instrument. Since eligibility is determined by employers, individual preferences for savings may play a minor role in the determination of eligibility after

controlling for observed covariates including family income, age, marital status and family size. Furthermore, it is plausible that 401(k) eligibility has an impact on participation in Individual Retirement Accounts only through participation in 401(k) plans. The monotonicity and instrumental variable relevance assumptions hold trivially, as only eligible individuals may choose to participate in 401(k) plans.

In our analysis, we use the dataset prepared for Abadie (2003), which contains 9275 individuals from the Survey of Income and Program Participation of 1991. The study participants were between 25 and 64 years of age, and had an annual income between \$10 000 and \$200 000 and a family size ranging from 1 to 13; 62.9% of them were married. Assumptions 1 and 2 of the instrumental variable model imply restrictions on the observed-data law (Pearl, 1995). Wang et al. (2017b, § 3) showed that these restrictions may be tested by applying a modified Gail–Simon test for interaction after recoding the data. Applying this test to the data considered by Abadie (2003) confirms that they are compatible with Assumptions 1 and 2 at α-level 0.05. The Gail–Simon test was performed conditional on the following discrete covariates: family income, with the categories under \$20 000, \$20 000–30 000, \$30 000–40 000, \$40 000–50 000, \$50 000–75 000, and above \$75 000; age, with the categories 29 or younger, 30–35, 36–44, 45–54, and 55 or older; and a marriage indicator.

In the following, we use the instrumental variable model of Abadie (2003) to estimate the multiplicative local average treatment effect of 401(k) participation on the probability of holding an Individual Retirement Account. The Supplementary Material gives a graphical representation of the instrumental variable model assumed in our analysis. Throughout we make the following assumption on the local average treatment effect:

$$\delta^{\mathrm{M}}(X) = \exp(\alpha^{\mathrm{T}} X), \tag{7}$$

where the covariates X include an intercept, family income, family income squared, age, marital status and family size. Since there are no defiers or always takers, the multiplicative local average treatment effect can also be interpreted as the multiplicative treatment effect of 401(k) participation among those who actually participated in 401(k) plans. We apply the following estimation methods evaluated in the simulations: mle, drw, drw.ogburn, dru.ogburn, dru.simple, ls.abadie and mle.crude. Since only eligible individuals may participate in 401(k) plans, one can show that $E(H \mid X) = E(Y \mid Z = 0, X)$ and $\phi_2(X) = 0$, where $H = H(Y, D, X)$. As in the proof of Theorem 1, one can show that in this situation, the models of $\mathrm{pr}(D = d, Y = y \mid Z = z, X)$ $(d, y, z \in \{0, 1\})$ can be determined by the models of $\delta^{\mathrm{M}}(X)$, $\phi_1(X)$, $\phi_3(X)$ and $\mathrm{OP}^{\mathrm{CO}}(X)$. We provide details of model specifications in the Supplementary Material. The confidence intervals are obtained based on 500 bootstrap samples.

Figure 2 compares coefficient estimates for model (7). For example, results from mle suggest that with each additional family member, the multiplicative effect of 401(k) participation on holding an Individual Retirement Account increases by $\exp(0.068) - 1 = 7.0\%$ (95% confidence interval $[-0.3\%, 15.1\%]$). Results for drw.ogburn are not plotted, as its variance is huge compared with the other estimators. This suggests that the model for $E(DY \mid Z = 1, X)$ or the model for $E[\{H - E(H \mid X)\}^2 / f^2(Z \mid X) \mid X]$, or both, is probably misspecified. For the rest, the 95% confidence intervals obtained using drw are narrower than those obtained using dru.ogburn and dru.simple. This suggests that adopting the optimal weighting function is useful for reducing the variability of effect estimates. None of the covariates considered here is a significant modifier for the crowding-out effect of the 401(k) plan at α-level 0.05.

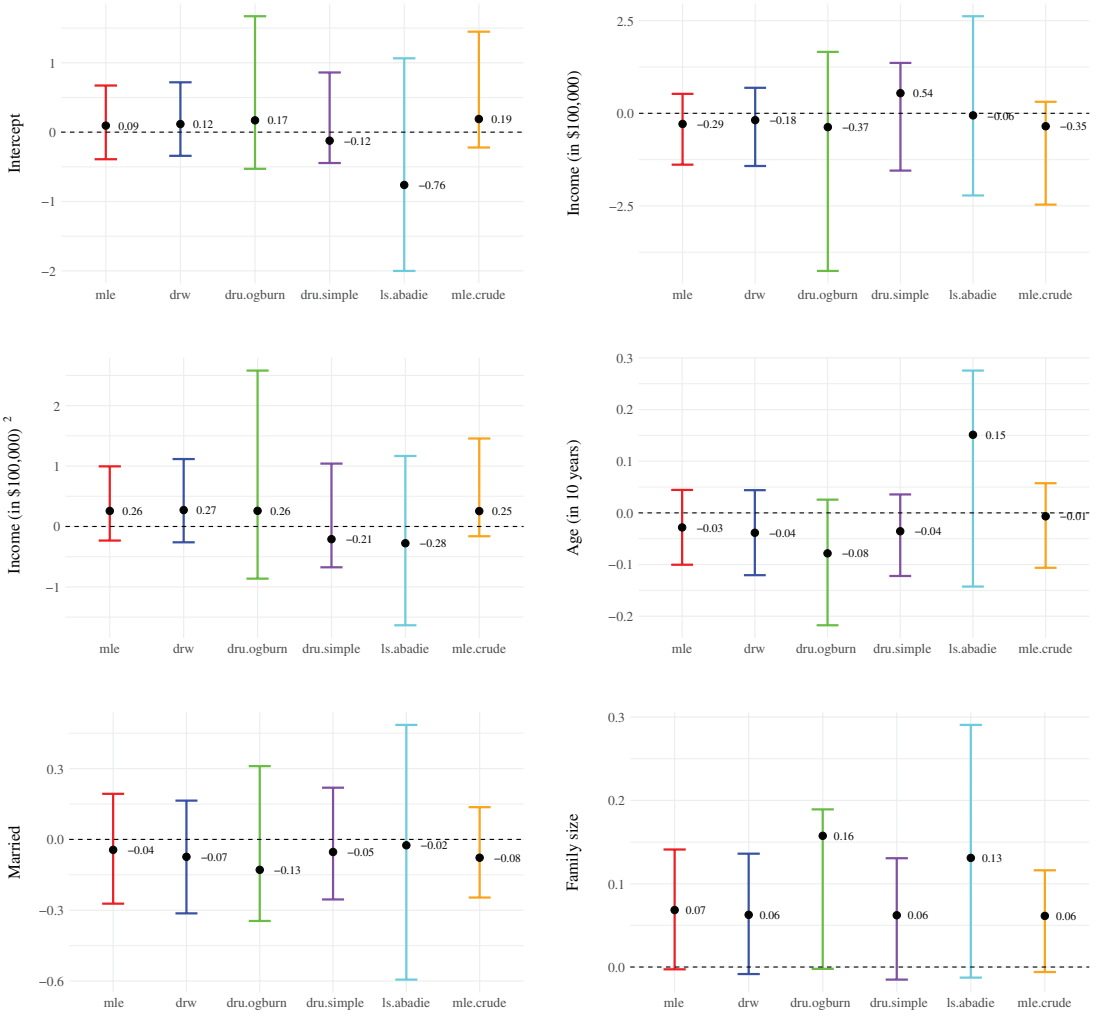

Fig. 2. Comparison of estimates of coefficients in the multiplicative local average treatment effect model (7) obtained using different methods. The black dots correspond to the point estimates, and the line segments represent the associated 95% confidence intervals.

We also examine representative subgroups of married and unmarried individuals and present the results in Fig. 3. The typical married subjects in this dataset, defined by the median of individual covariates, were 40 years old and had an annual income of $40 530 and a family of size of 4. Correspondingly, the typical unmarried subjects in this dataset were 39 years old and had an annual income of $23 718 and no other family members. Analysis results from mle suggest that for a typical married subject, participation in the 401(k) programme increases the likelihood of holding an Individual Retirement Account by 14.7% (95% confidence interval $[-1.2\%, 33.3\%]$). In comparison, participation in the 401(k) programme has virtually no effect on holding an Individual Retirement Account for a typical unmarried subject. In either case, there is no evidence for the crowding-out effect of 401(k) participation. Comparing the results of mle.crude with those of mle and drw, Fig. 3 shows that the instrumental variable methods attenuate the estimated effect of 401(k) participation on the probability of holding an Individual Retirement Account. These findings are consistent with the observations of Abadie (2003).

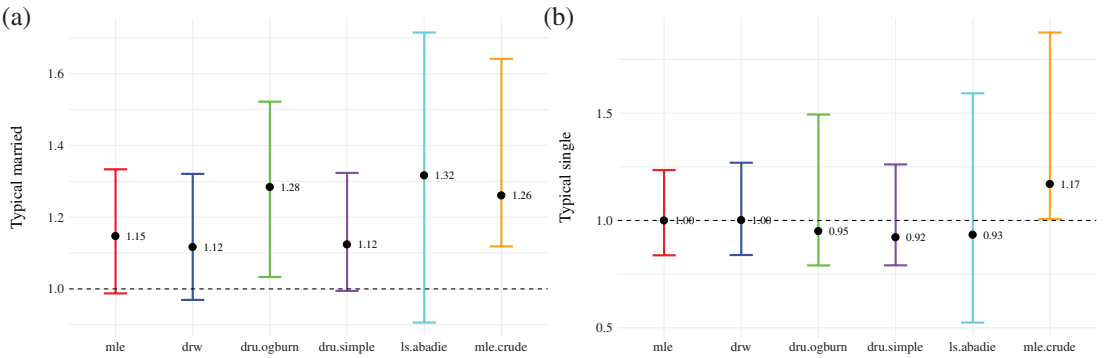

Fig. 3. Comparison of estimated multiplicative local average treatment effects within two representative subgroups. The black dots correspond to the point estimates of the local average treatment effect for (a) a typical married subject or (b) a typical unmarried subject. The line segments represent the associated 95% confidence intervals.

Acknowledgement

The authors thank Elizabeth Ogburn for helpful conversations, and the referees and associate editor for their insightful comments. This research was supported by the Natural Sciences and Engineering Research Council of Canada and the U.S. National Institutes of Health and Office of Naval Research.

Supplementary material

Supplementary Material available at *Biometrika* online contains additional details of the simulation studies and data analysis, as well as proofs of Theorems 1 and 2.

References

Abadie, A. (2002). Bootstrap tests for distributional treatment effects in instrumental variable models. *J. Am. Statist. Assoc.* **97**, 284–92.

Abadie, A. (2003). Semiparametric instrumental variable estimation of treatment response models. *J. Economet.* **113**, 231–63.

Abadie, A., Angrist, J. D. & Imbens, G. W. (2002). Instrumental variables estimates of the effect of subsidized training on the quantiles of trainee earnings. *Econometrica* **70**, 91–117.

Angrist, J. D., Imbens, G. W. & Rubin, D. B. (1996). Identification of causal effects using instrumental variables. *J. Am. Statist. Assoc.* **91**, 444–55.

Clarke, P. S. & Windmeijer, F. (2012). Instrumental variable estimators for binary outcomes. *J. Am. Statist. Assoc.* **107**, 1638–52.

Didelez, V., Meng, S. & Sheehan, N. A. (2010). Assumptions of IV methods for observational epidemiology. *Statist. Sci.* **25**, 22–40.

Frangakis, C. E. & Rubin, D. B. (2002). Principal stratification in causal inference. *Biometrics* **58**, 21–9.

Frölich, M. (2007). Nonparametric IV estimation of local average treatment effects with covariates. *J. Economet.* **139**, 35–75.

Goldberger, A. S. (1972). Structural equation methods in the social sciences. *Econometrica* **40**, 979–1001.

Hernán, M. A. & Robins, J. M. (2006). Instruments for causal inference: An epidemiologist's dream? *Epidemiology* **17**, 360–72.

Imbens, G. W. & Angrist, J. D. (1994). Identification and estimation of local average treatment effects. *Econometrica* **62**, 467–75.

Meng, X.-L. (1994). Multiple-imputation inferences with uncongenial sources of input. *Statist. Sci.* **9**, 538–73.

Ogburn, E. L., Rotnitzky, A. & Robins, J. M. (2015). Doubly robust estimation of the local average treatment effect curve. *J. R. Statist. Soc.* B **77**, 373–96.

Okui, R., Small, D. S., Tan, Z. & Robins, J. M. (2012). Doubly robust instrumental variable regression. *Statist. Sinica* **22**, 173–205.

PEARL, J. (1995). On the testability of causal models with latent and instrumental variables. In *Proc. 11th Conf. Uncertainty in Artificial Intelligence (UAI'95)*. San Francisco, California: Morgan Kaufmann Publishers, pp. 435–43.

PEARL, J. (2009). *Causality*. Cambridge: Cambridge University Press.

RICHARDSON, T. S., EVANS, R. J. & ROBINS, J. M. (2011). Transparent parameterizations of models for potential outcomes. *Bayesian Statist.* **9**, 569–610.

RICHARDSON, T. S. & ROBINS, J. M. (2013). Single world intervention graphs (SWIGs): A unification of the counterfactual and graphical approaches to causality. Working Paper no. 128, Center for the Statistics and the Social Sciences, University of Washington.

RICHARDSON, T. S., ROBINS, J. M. & WANG, L. (2017). On modeling and estimation for the relative risk and risk difference. *J. Am. Statist. Assoc.* **112**, 1121–30.

TAN, Z. (2006). Regression and weighting methods for causal inference using instrumental variables. *J. Am. Statist. Assoc.* **101**, 1607–18.

WANG, L., RICHARDSON, T. S. & ROBINS, J. M. (2017a). Congenial causal inference with binary structural nested mean models. *arXiv:* 1709.08281.

WANG, L., ROBINS, J. M. & RICHARDSON, T. S. (2017b). On falsification of the binary instrumental variable model. *Biometrika* **104**, 229–36.

WANG, L., ZHOU, X.-H. & RICHARDSON, T. S. (2017c). Identification and estimation of causal effects with outcomes truncated by death. *Biometrika* **104**, 597–612.

WANG, L. & TCHETGEN TCHETGEN, E. (2018). Bounded, efficient and multiply robust estimation of average treatment effects using instrumental variables. *J. R. Statist. Soc.* B **80**, 531–50.

WRIGHT, P. G. & WRIGHT, S. (1928). *The Tariff on Animal and Vegetable Oils*. New York: Macmillan.

[*Received on* 26 *April* 2019. *Editorial decision on* 29 *September* 2020]