Published in final edited form as:

Biometrics. 2023 June ; 79(2): 569–581. doi:10.1111/biom.13783.

Instrumented difference-in-differences

Ting Ye[1], **Ashkan Ertefaie**[2], **James Flory**[3], **Sean Hennessy**[4], **Dylan S. Small**[5]

[1]Department of Biostatistics, University of Washington, Seattle, Washington, USA

[2]Department of Biostatistics and Computational Biology, University of Rochester, Rochester, New York, USA

[3]Department of Subspecialty Medicine, Memorial Sloan Kettering Cancer Center, New York, New York, USA

[4]Department of Biostatistics, Epidemiology, and Informatics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania, USA

[5]Department of Statistics and Data Science, The Wharton School, University of Pennsylvania, Philadelphia, Pennsylvania, USA

Abstract

Unmeasured confounding is a key threat to reliable causal inference based on observational studies. Motivated from two powerful natural experiment devices, the instrumental variables and difference-in-differences, we propose a new method called instrumented difference-in-differences that explicitly leverages exogenous randomness in an exposure trend to estimate the average and conditional average treatment effect in the presence of unmeasured confounding. We develop the identification assumptions using the potential outcomes framework. We propose a Wald estimator and a class of multiply robust and efficient semiparametric estimators, with provable consistency and asymptotic normality. In addition, we extend the instrumented difference-in-differences to a two-sample design to facilitate investigations of delayed treatment effect and provide a measure of weak identification. We demonstrate our results in simulated and real datasets.

Keywords

causal inference; effect modification; exclusion restriction; instrumental variables; multiple robustness

Correspondence Ting Ye, Department of Biostatistics, University of Washington, Seattle, WA, USA. tingye1@uw.edu.

OPEN RESEARCH BADGES
This article has earned Open Data and Open Materials badges. Data and materials are available at https://github.com/jfiksel/compregpaper, https://github.com/jfiksel/codalm.

SUPPORTING INFORMATION
Web Appendices referenced in Sections 2–7 are available with this paper at the Biometrics website on Wiley Online Library. Code and data for replicating the simulations and data analyses in this paper are being made available online with the paper. R package idid, available at https://github.com/tye27/idid, implements the proposed method.
Table S1: Sample sizes for 1970 NHIS datasets and 1975, 1985, 1995, 2005 CDC WONDER compressed mortality datasets by birth cohort and gender
Data and codes are available in the Supporting Information of this article.

1 | INTRODUCTION

Unmeasured confounding is a key threat to reliable causal inference based on observational studies (Lawlor et al., 2004; Rutter, 2007). A popular approach to handle unmeasured confounding is the instrumental variable (IV) method, which requires an IV that satisfies three core assumptions (Angrist et al., 1996; Baiocchi et al., 2014; Hernan & Robins, 2020): (i) (relevance) it is associated with the exposure; (ii) (independence) it is independent of any unmeasured confounder of the exposure–outcome relationship; (iii) (exclusion restriction) it has no direct effect on the outcome. By extracting exogenous variation in the exposure that is independent of the unmeasured confounder, IVs can be used to estimate the causal effect.

Meanwhile, the increasing availability of large longitudinal datasets such as administrative claims and electronic health records has created new opportunities to expand study designs to take the advantage of the longitudinal structure. One method that is widely used in economics and other social sciences is difference-in-differences (DID) (Card & Krueger, 1994; Angrist & Pischke, 2008). The method of DID is based on a comparison of the trends in outcome for two exposure groups, where one group consists of individuals who switch from being unexposed to exposed and the other group consists of individuals who are never exposed. Under the parallel trends assumption, which says that the outcomes in the two exposure groups evolve in the same way over time in the absence of the exposure, DID is able to remove time-invariant bias from the unmeasured confounder. However, because the setup and assumptions of DID are motivated from applications in social sciences, its applicability is limited in biomedical sciences. For example, in social sciences it is relatively common for a new policy to be applied to one region of the country but not another, creating a circumstance in which key assumptions such as parallel trends are likely to hold and facilitating a DID design. In assignment of pharmacologic or other treatments in health care, such clear natural, exogenous sources of cleavage between exposed and unexposed groups are rare, making it more difficult to identify situations in which all assumptions of DID will be met.

In this paper, we connect these two powerful natural experiment devices (referred to as the standard IV and standard DID) and propose a new method called instrumented DID to estimate the causal effect of the exposure in the presence of unmeasured confounding. Unlike the standard DID, the instrumented DID exploits a *haphazard* encouragement targeted at a subpopulation toward faster uptake of the exposure or a surrogate of such encouragement, which we call *IV for DID*. Then, any observed nonparallel trends in outcome between the encouraged and unencouraged groups provides evidence for causation, as long as their trends in outcome are parallel if all individuals were counterfactually not exposed. A prototypical example of instrumented DID is a longitudinal randomized experiment, where after a baseline period, some individuals are randomly selected to be encouraged to take the treatment regardless of their treatment history. If the encouragement is effective, the exposure rate would increase more for the encouraged group than the unencouraged group. If additionally the encouragement has no direct effect on the trend in outcome, then any nonparallel trends in outcome must be due to the nonparallel trends in exposure. Therefore, through exploiting haphazard encouragement that affects the exposure trend, the instrumented DID is able to extract some variation in the exposure trend that is independent of the unmeasured confounder and relax some of the most disputable assumptions of the standard IV and standard DID method, particularly the exclusion restriction for the standard IV method and the parallel trends for the standard DID method; see Section 2 for more discussion.

Reasoning similar to the instrumented DID has been applied informally in prior studies. A prominent example is the differential trends in smoking prevalence for men and women as a consequence of targeted tobacco advertising to women, which were associated with disproportional trends for men and women in lung cancer mortality (Burbank, 1972; Meigs, 1977; Patel et al., 2004). Specifically, because of marketing efforts designed to introduce specific women's brands of cigarettes such as Virginia Slims in 1967, there was a considerable increase in smoking initiation by young women, which lasted through the mid-1970s (Pierce & Gilpin, 1995). Thirty years later, the lung cancer mortality rates for women at the age of 55 years or older had increased to almost four times the 1970 rate, whereas rates among men had no such dramatic change (Bailar & Gornik, 1997). In Section 7, we will analyze this example using the proposed method.

The rest of this paper is organized as follows. In Section 2, we establish the identification assumptions for the instrumented DID. In Section 3, we develop various estimation and inference approaches. In Section 4, we extend the instrumented DID to a two-sample design. In Section 5, we provide a measure of weak identification. Results from simulation studies and a real-data application are in Sections 6 and 7, respectively. The paper concludes with a discussion in Section 8. A review of IV and DID designs can be found in Section S1 of the Supporting information.

2 | INSTRUMENTED DIFFERENCE-IN-DIFFERENCES: IDENTIFICATION

Suppose that random samples of a target population are collected at two time points $t = 0$ and $t = 1$, and there is no overlap between individuals in these two samples. We leave consideration of overlapping samples and multiple time points to future work. For each individual i in the *pooled* sample, we observe $\boldsymbol{O}_i = (T_i, Z_i, \boldsymbol{X}_i, D_i, Y_i)$, where T_i is a time indicator that equals one if this individual appears at $t = 1$, equals zero if $t = 0$, Z_i is a binary IV for DID observed at the baseline, $\boldsymbol{X}_i$ is a vector of baseline covariates, D_i is a binary exposure variable, Y_i is a real-valued outcome of interest. We assume that $(\boldsymbol{O}_1, \ldots, \boldsymbol{O}_n)$ are independent and identically distributed (i.i.d.) realizations of $\boldsymbol{O} = (T, Z, \boldsymbol{X}, D, Y)$. This data setup is also commonly known as repeated cross-sectional data (Abadie, 2005).

We define causal effects using the potential outcomes framework (Neyman, 1923; Rubin, 1974). For each individual, let $D_t^{(z)}$ be the potential exposure if this individual were observed at time t and if Z were externally set to z, $Y_t^{(d)}$ be the potential outcome if this individual were observed and exposed to d at time t, and Z had the same value it actually had. The full data vector for each individual is $(Z, \boldsymbol{X}, D_t^{(z)}, Y_t^{(d)}, t = 0{,}1, z = 0{,}1, d = 0{,}1)$. Moreover, let $Y^{(d)}: = TY_1^{(d)} + (1 - T)Y_0^{(d)}$ be the potential outcome if this individual were exposed to d at the time point it actually got sampled and Z had the same value it actually had. Our target estimand is the average treatment effect (ATE) $\beta_0 = E(Y^{(1)} - Y^{(0)})$ and conditional average treatment effect (CATE) $\beta_0(\boldsymbol{v}) = E(Y^{(1)} - Y^{(0)} \mid \boldsymbol{V} = \boldsymbol{v})$, where $\boldsymbol{V}$ is a pre-specified subset of $\boldsymbol{X}$, representing the effect modifiers of interest; for example, setting $\boldsymbol{V}$ to be an empty set gives the unconditional ATE β_0. Note that the separation of $\boldsymbol{V}$ and $\boldsymbol{X}$ separates the need to adjust for possible confounding and the specification of effect modifiers of interest, which provides great flexibility and allows researchers to define the target estimand a priori. Throughout the paper, we consider treatment effect on the additive scale.

We make the following identification assumptions for using the instrumented DID.

Assumption 1.

a. (Consistency) $D = D_T^{(Z)}$ and $Y = Y_T^{(D)}$.

b. (Positivity) $0 < P(T = t, Z = z \mid \boldsymbol{X}) < 1$ for $t = 0,\ 1,\ z = 0,\ 1$ with probability 1.

c. (Random sampling) $T \perp (D_t^{(z)}, Y_t^{(d)}, t = 0,1, z = 0,\ 1, d = 0,1) \mid Z, \boldsymbol{X}$.

Assumption 1(a) states that the observed exposure is $D = D_t^{(z)}$ if and only if $Z = z$ and $T = t$, and the observed outcome is $Y = Y_t^{(d)}$ if and only if $D = d$ and $T = t$. Implicit in this assumption is that an individual's observed outcome is not affected by others' exposure level or this individual's exposure level at the other time point; this is known as the Stable Unit Treatment Value Assumption (Rubin, 1978, 1990). Assumption 1(b) postulates that there is a positive probability of receiving each (t, z) combination within each level of $\boldsymbol{X}$, or equivalently, the support of $\boldsymbol{X}$ is the same for each level of (T, Z). Assumption 1(c) is often assumed for repeated cross-sectional studies and says that for each level of $(Z, \boldsymbol{X})$, the collected data at every time point are a random sample from the underlying population; see Section 3.2.1 of Abadie (2005) that makes a similar assumption.

Assumption 2.

(Instrumented DID). With probability 1,

a. (Trend relevance) $E(D_1^{(1)} - D_0^{(1)} \mid Z = 1, \boldsymbol{X}) \neq E(D_1^{(0)} - D_0^{(0)} \mid Z = 0, \boldsymbol{X})$.

b. (Independence & exclusion restriction) $Z \perp (D_t^{(0)}, D_t^{(1)}, Y_1^{(0)} - Y_0^{(0)}, Y_t^{(1)} - Y_t^{(0)}, t = 0,1) \mid \boldsymbol{X}$.

c. (No unmeasured common effect modifier) $Cov(D_t^{(1)} - D_t^{(0)}, Y_t^{(1)} - Y_t^{(0)} \mid \boldsymbol{X}) = 0$ for $t = 0,\ 1$.

d. (Stable treatment effect over time) $E(Y_1^{(1)} - Y_1^{(0)} \mid \boldsymbol{X}) = E(Y_0^{(1)} - Y_0^{(0)} \mid \boldsymbol{X})$.

Assumptions 2(a) and (b) formalize the core assumptions that an IV for DID needs to satisfy, which are illustrated by a directed acyclic graph (DAG) in Figure 1. Assumptions 2(a) and (b) are also parallel to the core assumptions for the standard IV introduced in Section 1.

Assumption 2(a) says that the IV for DID Z, as an encouragement that disproportionately acts on only a subpopulation, affects the trend in exposure. For example, Z can be a random encouragement for some subjects in a longitudinal experiment, an advertisement campaign targeted at a certain geographic region or subpopulation, or a change in reimbursement policies for a certain insurance plan. Under Assumption 1, Assumption 2(a) is equivalent to $E(D \mid T = 1, Z = 1, \boldsymbol{X}) - E(D \mid T = 0, Z = 1, \boldsymbol{X}) \neq E(D \mid T = 1, Z = 0, \boldsymbol{X}) - E(D \mid T = 0, Z = 0, \boldsymbol{X})$ with probability 1, thus is checkable from observed data.

Assumption 2(b) is an integration of the independence and exclusion restriction assumptions. To see this, we adopt a more elaborated definition of the potential outcomes and define $Y_t^{(dz)}$ as the potential outcome if the individual were observed and exposed to d at time t, and if Z were externally set to z, then Assumption 2(b) is implied by (independence) $Z \perp (D_t^{(0)}, D_t^{(1)}, Y_1^{(0z)} - Y_0^{(0z)}, Y_t^{(1z)} - Y_t^{(0z)}, t = 0{,}1, z = 0{,}1) \mid \boldsymbol{X}$ and (exclusion restriction) $Y_t^{(11)} - Y_t^{(01)} \mid \boldsymbol{X} \sim_d Y_t^{(10)} - Y_t^{(00)} \mid \boldsymbol{X}$ and $Y_1^{(01)} - Y_0^{(01)} \mid \boldsymbol{X} \sim_d Y_1^{(00)} - Y_0^{(00)} \mid \boldsymbol{X}$, where $\sim_d$ means having the same distribution; see Tan (2006) for a parallel statement for the standard IV and Hernán and Robins (2006) for connections and comparisons between different definitions of the standard IV. Hence, Assumption 2(b) essentially states that Z is unconfounded, has no direct effect on the trend in outcome, and does not modify the treatment effect. Here, we see the main advantage of using Z as an IV for DID compared to as a standard IV: Z as an IV for DID is allowed to have a direct effect on the outcome, as long as it has no direct effect on the trend in outcome and does not modify the treatment effect. For example, Newman et al. (2012) considered using a hospital's preference for phototherapy when treating newborns with hyperbilirubinemia as a standard IV to study the effect of phototherapy but found evidence that hospitals that use more phototherapy also have greater use of infant formula, which is thought to be an effective treatment for hyperbilirubinemia. Hence, the hospital's preference is a potentially invalid standard IV as it can have a direct effect on the outcome through the use of infant formula. However, it may still qualify as an IV for DID if the use of phototherapy evolves differently between the high and low preference hospitals over time, but the use of infant formula in the two groups of hospitals does not change over time. These features imply that variables like hospital's preference may be more likely to be an IV for DID, compared to being a standard IV.

Assumption 2(c) is developed in Cui and Tchetgen Tchetgen (2021) and a slightly stronger version is proposed earlier in Wang and Tchetgen Tchetgen (2018). Suppose in this paragraph only the existence of an unmeasured confounder U_t such that $(D_t^{(1)}, D_t^{(0)}) \perp (Y_t^{(1)}, Y_t^{(0)}) \mid (U_t, \boldsymbol{X})$, then Assumption 2(c) holds if either (i) there is no additive $U_t - z$ interaction in $E(D_t^{(z)} \mid U_t, \boldsymbol{X})$: $E(D_t^{(1)} - D_t^{(0)} \mid U_t, \boldsymbol{X}) = E(D_t^{(1)} - D_t^{(0)} \mid \boldsymbol{X})$; or (ii) there is no additive $U_t - d$ interaction in $E(Y_t^{(d)} \mid U_t, \boldsymbol{X})$: $E(Y_t^{(1)} - Y_t^{(0)} \mid U_t, \boldsymbol{X}) = E(Y_t^{(1)} - Y_t^{(0)} \mid \boldsymbol{X})$.

Assumption 2(d) requires that the CATE does not change over time. This is a strong assumption but may be plausible in many applications when the study period only spans a short period of time. In our application in Section 7, we conduct a sensitivity analysis to gauge the sensitivity of the study conclusion to violation of this assumption.

Two additional remarks on Assumption 2 are in order. First, an attractive feature of Assumptions 2(c) and (d) is that they are guaranteed to be true under the sharp null hypothesis of no treatment effect for all individuals. This means that the instrumented DID method can be used for testing the sharp null hypothesis under Assumptions 2(a) and (b). Second, from the definition of potential exposures $D_t^{(z)}$, the IV for DID Z is considered causal for the exposure. In Section S4.2, we present another version of notations and assumptions which does not require Z to be causal, that is, Z is allowed to be correlated with a cause that affects the trend in exposure, and is more suitable for our application in Section 7 in which we use gender as the IV for DID for its correlation with the encouragement from targeted tobacco advertising.

For $C \in \{Y, D\}$, let $\mu_C(t, z, \boldsymbol{X}) = E(C \mid T = t, Z = , \boldsymbol{X})$, $\delta_C(\boldsymbol{X}) = \mu_C(1,1, \boldsymbol{X}) - \mu_C(0,1, \boldsymbol{X}) - \mu_C(1,0, \boldsymbol{X}) + \mu_C(0,0, \boldsymbol{X})$, and let $\mu_C(t, z)$ and δ_C denote their counterparts without observed covariates. The next proposition indicates that the (conditional) ATE can be identified under the above assumptions.

Proposition 1.

If Assumptions 1 and 2 hold, then

$$\delta(\boldsymbol{X}) := \frac{\delta_Y(\boldsymbol{X})}{\delta_D(\boldsymbol{X})} = \beta_0(\boldsymbol{X}) \textit{ and } E[\delta(\boldsymbol{X}) \mid \boldsymbol{V} = \boldsymbol{v}] = \beta_0(\boldsymbol{v}) . \tag{1}$$

This and all the other proofs in this paper are in Section S3.

Now we contrast the instrumented DID with standard DID. As discussed in Section 1, the standard DID identifies the ATE for the treated in the post-treatment period from comparing the trends in outcome between two exposure groups, where every individual in one group switches from being unexposed to exposed between two time points, and every individual in the other group is never exposed. However, its key assumption, the parallel trends, will be violated if there is a time-invariant unmeasured confounder that has time-varying effects or there is a time-varying unmeasured confounder in the exposure–outcome relationship. We use time-varying unmeasured confounding to refer to either case. In contrast, the instrumented DID explicitly probes the relationship between the trend in outcome and the trend in exposure using an exogenous variable Z which often results in partial compliance with exposure within groups defined by levels of Z. Compared with standard DID, instrumented DID is robust to time-varying unmeasured confounding in the exposure–outcome relationship by making use of an exogenous variable Z that is not subject to this time-varying unmeasured confounding.

We remark that when there are no observed covariates, δ_Y / δ_D has been derived in alternative ways in econometrics under different assumptions. It is the same as the standard IV Wald ratio after first differencing the exposure and outcome when each individual is observed at both time points (Wooldridge, 2010, Chap. 15.8), as motivated from the linear structural equation models. Importantly, Proposition 1 provides a justification of this approach using the potential outcomes framework without any modeling assumption. It is also the same as the Wald ratio in the fuzzy DID method for identification of a local ATE under the assumption that individuals can switch treatment in only one direction within each treatment group (de Chaisemartin & D’HaultfŒuille, 2018), as motivated from social science applications (e.g., Duflo 2001). Compared with this derivation, our proposed instrumented DID is less stringent in terms of the direction in which each individual can switch treatment, thus is better suited for applications using healthcare data where individuals can switch treatment in any direction. In addition, we complement the proposed instrumented DID with a novel semiparametric estimation and inference method in Section 3.2, two-sample design in Section 4, and measure of weak identification in Section 5.

Finally, we note that Assumption 2(c) can be replaced by the monotonicity assumption $D_t^{(1)} \geq D_t^{(0)}$ for $t = 0,\ 1$ with probability 1, under which $\delta(\boldsymbol{X})$ in (1) identifies a complier ATE; see Section S3.3 for details.

3 | ESTIMATION AND INFERENCE

3.1 | Wald estimator

When there are no observed covariates, based on Proposition 1, we can simply replace the conditional expectations in Equation (1) with their sample analogues and obtain the Wald estimator

$$\widehat{\beta}_{\text{wald}} = \frac{\widehat{\mu}_Y(1,1) - \widehat{\mu}_Y(0,1) - \widehat{\mu}_Y(1,0) + \widehat{\mu}_Y(0,0)}{\widehat{\mu}_D(1,1) - \widehat{\mu}_D(0,1) - \widehat{\mu}_D(1,0) + \widehat{\mu}_D(0,0)} = \frac{\widehat{\delta}_Y}{\widehat{\delta}_D}, \tag{2}$$

where $\widehat{\mu}_C(t, z) = \sum_{i=1}^n C_i I_{(T_i = t, Z_i = z)} / \sum_{i=1}^n I_{(T_i = t, Z_i = z)}$, $\widehat{\delta}_C = \widehat{\mu}_C(1,1) - \widehat{\mu}_C(0,1) - \widehat{\mu}_C(1,0) + \widehat{\mu}_C(0,0)$, for $C \in \{Y, D\}$. In Theorem S1, we prove consistency and asymptotic normality of $\widehat{\beta}_{\text{wald}}$ and give a consistent variance estimator.

3.2 | Semiparametric theory and multiply robust estimators

Consider the case with a baseline observed covariate vector $\boldsymbol{X}$. Suppose that we have a parametric model for $\beta_0(\boldsymbol{v})$, written as $\beta(\boldsymbol{v}; \boldsymbol{\psi})$ for some finite-dimensional parameter $\boldsymbol{\psi}$. Importantly, we do not assume that this model is necessarily correct, but instead treat it as a working model and formulate our estimand as the projection of the CATE $\beta_0(\boldsymbol{v})$ onto the *working model* $\beta(\boldsymbol{v}; \boldsymbol{\psi})$. Specifically, we use the weighted least-squares projection given by

$$\boldsymbol{\psi}_0 = \underset{\boldsymbol{\psi}}{\operatorname{argmin}}\ E\left[w(\boldsymbol{V})\{\beta_0(\boldsymbol{V}) - \beta(\boldsymbol{V}; \boldsymbol{\psi})\}^2\right], \tag{3}$$

where $w(\boldsymbol{v})$ is a user-specified weight function, which can be tailored if there is subject matter knowledge for emphasizing specific parts of the support of $\boldsymbol{V}$; otherwise, we can set $w(\boldsymbol{v}) = 1$. By definition, $\beta(\boldsymbol{V}; \boldsymbol{\psi}_0)$ is the best least-squares approximation to the CATE $\beta_0(\boldsymbol{V})$. For example, when effect modification is not of interest, we can specify $\beta(\boldsymbol{v}; \boldsymbol{\psi}) = \boldsymbol{\psi}$ such that $\beta_0(\boldsymbol{V})$ is projected onto a constant $\boldsymbol{\psi}_0$, which can be interpreted as the ATE; if we want to estimate a linear approximation of the CATE, we can specify $\beta\left(\boldsymbol{v}; \boldsymbol{\psi}\right) = \boldsymbol{v}^T \boldsymbol{\psi}$, with $\boldsymbol{V}$ including the intercept. This working model approach is also adopted in Abadie (2003), Ogburn et al. (2015), and Kennedy et al. (2019).

Let $\pi(t, z, \boldsymbol{x}) = P(T = t, Z = z \mid \boldsymbol{X} = \boldsymbol{x})$, $b_C(\boldsymbol{x}) = \mu_C(0,0, \boldsymbol{x})$, $m_{CZ}(\boldsymbol{x}) = \mu_C(0,1, \boldsymbol{x}) - \mu_C(0,0, \boldsymbol{x})$, $m_{CT}(\boldsymbol{x}) = \mu_C(1,0, \boldsymbol{x}) - \mu_C(0,0, \boldsymbol{x})$, and $\boldsymbol{\Delta}_C(\boldsymbol{x}) = (b_C(\boldsymbol{x}), m_{CZ}(\boldsymbol{x}), m_{CT}(\boldsymbol{x}))$, for $C \in \{Y, D\}$. Consider three sets of model assumptions:

$\mathscr{M}_1$: models for $\delta(\boldsymbol{x}), \boldsymbol{\Delta}_D(\boldsymbol{x}), \boldsymbol{\Delta}_Y(\boldsymbol{x})$ are correct.

$\mathscr{M}_2$: models for $\pi(t, z, \boldsymbol{x})$, $\delta_D(\boldsymbol{x})$ are correct.

$\mathscr{M}_3$: models for $\pi(t, z, \boldsymbol{x})$, $\delta(\boldsymbol{x})$ are correct.

In what follows, we first discuss three different estimators for $\boldsymbol{\psi}$ that are consistent and asymptotically normal under $\mathscr{M}_1$, $\mathscr{M}_2$, $\mathscr{M}_3$, respectively. Bounded semiparametric estimators analogous to those in Wang and Tchetgen Tchetgen (2018) are developed in Section S2.4. Let $\mathbb{P}_n\boldsymbol{X} = n^{-1}\sum_{i=1}^{n} \boldsymbol{X}_i$ be the empirical average. Under model $\mathscr{M}_1$, we present a regression-based estimator $\widehat{\boldsymbol{\psi}}_{\text{reg}}$ that solves

$$\mathbb{P}_n q(\boldsymbol{V}; \boldsymbol{\psi})\{\delta(\boldsymbol{x}; \widehat{\boldsymbol{\alpha}}_{\text{reg}}) - \beta(\boldsymbol{V}; \boldsymbol{\psi})\} = 0,$$

where $q(\boldsymbol{v}; \boldsymbol{\psi}) = w(\boldsymbol{v})\partial\beta(\boldsymbol{v}; \boldsymbol{\psi})/\partial\boldsymbol{\psi}, \delta(\boldsymbol{x}; \boldsymbol{\alpha})$ is a parametric specification of $\delta(\boldsymbol{x})$, $\widehat{\boldsymbol{\alpha}}_{\text{reg}}$ the solution to $\mathbb{P}_n h_\alpha(\boldsymbol{X})$ $\{Y - \widehat{b}_Y(\boldsymbol{X}) - \hat{m}_{YZ}(\boldsymbol{X})Z - \hat{m}_{YT}(\boldsymbol{X})T - \delta(\boldsymbol{X}; \boldsymbol{\alpha})(D - \widehat{b}_D(\boldsymbol{X}) - \widehat{m}_{DZ}(\boldsymbol{X})Z - \widehat{m}_{DT}(\boldsymbol{X})T)\} = 0,\ h_\alpha(\boldsymbol{X})$ a vector of the same dimension as $\boldsymbol{\alpha}$, and $\widehat{b}_D$, $\hat{m}_{DZ}$, $\hat{m}_{DT}$, $\widehat{b}_Y$, $\hat{m}_{YZ}$, $\hat{m}_{YT}$ are respectively estimators of b_D, m_{DZ}, m_{DT}, b_Y, m_{YZ}, m_{YT}. Under model $\mathscr{M}_2$, we present an inverse probability weighting (IPW) estimator $\widehat{\boldsymbol{\psi}}_{\text{ipw}}$ that solves

$$\mathbb{P}_n q(\boldsymbol{V}; \boldsymbol{\psi})\left\{\frac{(2Z-1)(2T-1)Y}{\widehat{\pi}(T, Z, \boldsymbol{X})\delta_D(\boldsymbol{X}; \widehat{\boldsymbol{\theta}})} - \beta(\boldsymbol{V}; \boldsymbol{\psi})\right\} = 0,$$

where $\delta_D(\boldsymbol{x}; \boldsymbol{\theta})$ is a parametric specification of $\delta_D(\boldsymbol{x})$, $\hat{\boldsymbol{\theta}}$ the solution to $\mathbb{P}_n h_\theta(\boldsymbol{X})\{(2Z-1)(2T-1)D/\widehat{\pi}(T, Z, \boldsymbol{X}) - \delta_D(\boldsymbol{X}; \theta)\} = 0, \widehat{\pi}(t, z, \boldsymbol{x})$ an estimator of $\pi(t, z, \boldsymbol{x})$, and $h_\theta(\boldsymbol{X})$ a vector of the same dimension as $\boldsymbol{\theta}$. Finally, under model $\mathscr{M}_3$, we present an estimator $\widehat{\boldsymbol{\psi}_{\text{g}}}$ based on g-estimation, defined as the solution to

$$\mathbb{P}_n q(\boldsymbol{V}; \boldsymbol{\psi})\{\delta(\boldsymbol{x}; \widehat{\boldsymbol{\alpha}}_{\text{g}}) - \beta(\boldsymbol{V}; \boldsymbol{\psi})\} = 0,$$

where $\widehat{\boldsymbol{\alpha}}_{\text{g}}$ is the solution to $\mathbb{P}_n h_\alpha(\boldsymbol{X})\{(2Z-1)(2T-1)(Y - \delta(\boldsymbol{X}; \boldsymbol{\alpha})D)/\hat{\pi}(T, Z, \boldsymbol{X})\} = 0$. These three classes of estimators are consistent and asymptotically normal in three different models $\mathscr{M}_1$, $\mathscr{M}_2$, $\mathscr{M}_3$, following standard arguments, for example, as in Newey and McFadden (1994, Chap. 6.1). Depending on the specific applications, some classes may be more preferable when knowledge about certain nuisance parameters is available. In practice, when we are uncertain about which models are correctly specified, it is of interest to develop a multiply robust estimator that is guaranteed to deliver valid inference about $\boldsymbol{\psi}_0$ provided that one, but not necessarily more than one, of models $\mathscr{M}_1$, $\mathscr{M}_2$, $\mathscr{M}_3$ holds (Vansteelandt et al., 2008; Wang & Tchetgen Tchetgen, 2018; Shi et al., 2020).

The next theorem derives the efficient influence function for $\boldsymbol{\psi}$ (Bickel et al., 1993; van der Vaart, 2000), which provides the basis of constructing a multiply robust estimator.

Theorem 1.—*If* Assumptions 1 *and* 2 *hold, and* $\partial\beta(\boldsymbol{v};\boldsymbol{\psi})/\partial\boldsymbol{\psi}$ *exists and is continuous. Under a nonparametric model, the efficient influence function for* $\boldsymbol{\psi}$ *is proportional to*

$$\varphi(\boldsymbol{O};\boldsymbol{\psi},\boldsymbol{\eta}) = q(\boldsymbol{V};\boldsymbol{\psi})(\delta(\boldsymbol{X}) - \beta(\boldsymbol{V};\boldsymbol{\psi}) + \frac{(2Z-1)(2T-1)}{\pi(T,Z,\boldsymbol{X})\delta_D(\boldsymbol{X})}$$

$$\begin{aligned}&[Y - b_Y(\boldsymbol{X}) - m_{YZ}(\boldsymbol{X})Z - m_{YT}(\boldsymbol{X})T - \delta(\boldsymbol{X})\{D - b_D(\boldsymbol{X})\\ &-m_{DZ}(\boldsymbol{X})Z - m_{DT}(\boldsymbol{X})T\}]),\end{aligned} \tag{4}$$

where $\boldsymbol{\eta} = (\pi,\delta_D,\delta,\boldsymbol{\Delta}_D,\boldsymbol{\Delta}_Y)$ *denotes the vector of nuisance parameters,* $\boldsymbol{\Delta}_D = (b_D, m_{DZ}, m_{DT})$, $\boldsymbol{\Delta}_Y = (b_Y, m_{YZ}, m_{YT})$, *and* $q(\boldsymbol{v};\boldsymbol{\psi}) = w(\boldsymbol{v})\partial\beta(\boldsymbol{v};\boldsymbol{\psi})/\partial\boldsymbol{\psi}$.

Note that the efficient influence function gives an estimator $\widehat{\boldsymbol{\psi}}_{\mathrm{mr}}$ defined as a solution to $\mathbb{P}_n\varphi(\boldsymbol{O};\boldsymbol{\psi},\hat{\boldsymbol{\eta}}) = 0$, where $\hat{\boldsymbol{\eta}}$ is a vector of the estimated nuisance parameters. Among the nuisance parameters, π, $\boldsymbol{\Delta}_D$, $\boldsymbol{\Delta}_Y$ can be estimated directly from likelihood or moment equations, whereas the estimation of δ_D and δ relies on additional nuisance parameters. To achieve multiple robustness, we need to construct a consistent estimator of δ_D in the union of $\mathscr{M}_1$ and $\mathscr{M}_2$, as well as a consistent estimator of δ in the union of $\mathscr{M}_1$ and $\mathscr{M}_3$. We achieve these goals by using doubly robust g-estimation (Robins, 1994). Specifically, we solve for $\widehat{\delta}_D(\boldsymbol{x}) = \delta_D(\boldsymbol{x};\widehat{\boldsymbol{\theta}}_{\mathrm{dr}})$ and $\widehat{\delta}(\boldsymbol{x}) = \delta(\boldsymbol{x};\widehat{\boldsymbol{\alpha}}_{\mathrm{dr}})$ respectively from

$$\begin{aligned}&\mathbb{P}_n h_\theta(\boldsymbol{X})\bigg[\frac{(2Z-1)(2T-1)}{\hat{\pi}(T,Z,\boldsymbol{X})}\Big\{D - \widehat{b}_D(\boldsymbol{X}) - \widehat{m}_{DZ}(\boldsymbol{X})Z\\ &-\widehat{m}_{DT}(\boldsymbol{X})T - \delta_D(\boldsymbol{X};\theta)ZT\}] = 0,\end{aligned}$$

$$\begin{aligned}&\mathbb{P}_n h_\alpha(\boldsymbol{X})\bigg[\frac{(2Z-1)(2T-1)}{\widehat{\pi}(T,Z,\boldsymbol{X})}\Big\{Y - \widehat{b}_Y(\boldsymbol{X}) - \widehat{m}_{YZ}(\boldsymbol{X})Z\\ &-\hat{m}_{YT}(\boldsymbol{X})T - \delta(\boldsymbol{X};\boldsymbol{\alpha})(D - \widehat{b}_D(\boldsymbol{X}) - \hat{m}_{DZ}(\boldsymbol{X})Z - \hat{m}_{DT}(\boldsymbol{X})T)\Big\}\bigg] = 0\,.\end{aligned}$$

We prove in the Supporting information that $\widehat{\boldsymbol{\psi}}_{\mathrm{mr}}$ is multiply robust, in the sense that the estimator is consistent as long as either one of the three models $(\mathscr{M}_1, \mathscr{M}_2, \mathscr{M}_3)$ holds.

Next, we derive the asymptotic properties of $\widehat{\boldsymbol{\psi}}_{\mathrm{mr}}$. Let $\xrightarrow{p}$ denote convergence in probability, $\parallel \boldsymbol{\psi} \parallel\ = \left(\boldsymbol{\psi}^T\boldsymbol{\psi}\right)^{1/2}$ the Euclidean norm for any column vector $\boldsymbol{\psi}$, $\parallel f \parallel_2\ = \left\{\int f^2(o)d\boldsymbol{P}(o)\right\}^{1/2}$ the $L_2(\boldsymbol{P})$ norm for any real-valued function f, and $\parallel \boldsymbol{f} \parallel_2\ = \sum_{j=1}^{\ell} \parallel f_j \parallel_2$ for any collection of real-valued functions $\boldsymbol{f} = (f_1,\ldots,f_\ell)$, where $\boldsymbol{P}$ denotes the distribution of $\boldsymbol{O}$. Moreover, let $\boldsymbol{\eta}_0 = (\pi_0,\delta_{D0},\delta_0,\boldsymbol{\Delta}_{D0},\boldsymbol{\Delta}_{Y0})$ denote the true values of the nuisance parameters.

Assumption 3.

a. $(\widehat{\boldsymbol{\psi}}_{\mathrm{mr}}, \hat{\boldsymbol{\eta}}) \xrightarrow{p} (\boldsymbol{\psi}_0, \overline{\boldsymbol{\eta}})$, where $\overline{\boldsymbol{\eta}} = \left(\overline{\pi}, \overline{\delta}_D, \overline{\delta}, \overline{\boldsymbol{\Delta}}_D, \overline{\boldsymbol{\Delta}}_Y\right)$ with either (i) $\overline{\delta} = \delta_0$, $\overline{\boldsymbol{\Delta}}_D = \boldsymbol{\Delta}_{D0}$, $\overline{\boldsymbol{\Delta}}_Y = \boldsymbol{\Delta}_{Y0}$; or (ii) $\overline{\pi} = \pi_0$ and $\overline{\delta}_D = \delta_{D0}$; or (iii) $\overline{\pi} = \pi_0$ and $\overline{\delta} = \delta_0$.

b. For each $\boldsymbol{\psi}$ in an open subset of Euclidean space and each $\boldsymbol{\eta}$ in a metric space, let $\varphi(\boldsymbol{o}; \boldsymbol{\psi}, \boldsymbol{\eta})$ be a measurable function such that the class of functions $\{\varphi(\boldsymbol{o}; \boldsymbol{\psi}, \boldsymbol{\eta}): \parallel \boldsymbol{\psi} - \boldsymbol{\psi}_0 \parallel < \epsilon, \parallel \boldsymbol{\eta} - \overline{\boldsymbol{\eta}} \parallel_2 < \epsilon\}$ is Donsker for some $\epsilon > 0$, and such that $E\parallel \varphi(\boldsymbol{O}; \boldsymbol{\psi}, \boldsymbol{\eta}) - \varphi(\boldsymbol{O}; \boldsymbol{\psi}_0, \overline{\boldsymbol{\eta}}) \parallel^2 \to 0$ as $(\boldsymbol{\psi}, \boldsymbol{\eta}) \to (\boldsymbol{\psi}_0, \overline{\boldsymbol{\eta}})$. The maps $\boldsymbol{\psi} \mapsto E\{\varphi(\boldsymbol{O}; \boldsymbol{\psi}, \boldsymbol{\eta})\}$ are differentiable at $\boldsymbol{\psi}_0$, uniformly in $\boldsymbol{\eta}$ in a neighborhood of $\overline{\boldsymbol{\eta}}$ with nonsingular derivative matrices $\boldsymbol{M}_{\psi_0, \eta} \to \boldsymbol{M}_{\psi_0, \overline{\eta}}$.

Assumption 3(a) describes the multiple robustness of our estimator. Assumption 3(b) is standard for M-estimators (van der Vaart, 2000, Chap. 5.4).

Theorem 2.—*Under* Assumptions 1–3, $\widehat{\boldsymbol{\psi}}_{\text{mr}}$ *is consistent with rate of convergence*

$$\parallel \widehat{\boldsymbol{\psi}}_{\text{mr}} - \boldsymbol{\psi}_0 \parallel = O_p\Big(n^{-1/2} + \parallel \hat{\delta} - \delta_0 \parallel_2 (\parallel \widehat{\pi} - \pi_0 \parallel_2 + \parallel \hat{\delta}_D \\ - \delta_{D0} \parallel_2) + \parallel \widehat{\pi} - \pi_0 \parallel_2 (\parallel \widehat{\boldsymbol{\Delta}}_Y - \boldsymbol{\Delta}_{Y0} \parallel_2 + \parallel \widehat{\boldsymbol{\Delta}}_D - \boldsymbol{\Delta}_{D0} \parallel_2) .$$

Suppose further that

$$\parallel \hat{\delta} - \delta_0 \parallel_2 (\parallel \widehat{\pi} - \pi_0 \parallel_2 + \parallel \hat{\delta}_D - \delta_{D0} \parallel_2) + \parallel \widehat{\pi} - \pi_0 \parallel_2 \\ (\parallel \widehat{\boldsymbol{\Delta}}_Y - \Delta_{Y0} \parallel_2 + \parallel \widehat{\boldsymbol{\Delta}}_D - \boldsymbol{\Delta}_{D0} \parallel_2 = o_p\Big(n^{-1/2}\Big),$$

then $\widehat{\boldsymbol{\psi}}_{\text{mr}}$ *is asymptotically normal and semiparametric efficient, satisfying*

$$\sqrt{n}(\widehat{\boldsymbol{\psi}}_{\text{mr}} - \boldsymbol{\psi}_0) \overset{d}{\to} N\Big(0, \boldsymbol{M}_{\psi_0, \eta_0}^{-1} E\{\varphi(\boldsymbol{O}; \boldsymbol{\psi}_0, \boldsymbol{\eta}_0) \\ \varphi(\boldsymbol{O}; \boldsymbol{\psi}_0, \boldsymbol{\eta}_0)^T\Big\}(\boldsymbol{M}_{\psi_0, \eta_0}^{-1})^T\Big) . \qquad (5)$$

The first part of Theorem 2 describes the convergence rate of $\widehat{\boldsymbol{\psi}}_{\text{mr}}$, which again indicates the multiple robustness of our estimator. That is, $\widehat{\boldsymbol{\psi}}_{\text{mr}}$ is consistent provided that (i) either one of $\hat{\pi}$ or $(\widehat{\boldsymbol{\Delta}}_Y, \widehat{\boldsymbol{\Delta}}_D)$ is consistent, and (ii) either one of $\hat{\delta}$ or $(\hat{\pi}, \hat{\delta}_D)$ is consistent. The multiple robustness property is important in practice, because nuisance parameters such as π, δ_D, and δ may be easier to estimate than $\boldsymbol{\Delta}_Y$ and $\boldsymbol{\Delta}_D$. When all the nuisance parameters are consistently estimated, we can still benefit from using the semiparametric methods, in that even the nuisance parameters are estimated at slower rates, $\widehat{\boldsymbol{\psi}}_{\text{mr}}$ can still have the fast convergence rate. For example, if all the nuisance parameters are estimated at $n^{-1/4}$ rates, then $\widehat{\boldsymbol{\psi}}_{\text{mr}}$ can still achieve fast $n^{-1/2}$ rate. The second part of Theorem 2 says that if the nuisance parameters are consistently estimated with fast rates, for example, if they are estimated using parametric methods, then their variance contributions are negligible, and $\widehat{\boldsymbol{\psi}}_{\text{mr}}$ achieves the semiparametric efficiency bound.

When Equation (5) holds, a plug-in variance estimator for $\sqrt{n}\widehat{\boldsymbol{\psi}}_{\text{mr}}$ can be easily constructed as $\widehat{\boldsymbol{M}}^{-1}\{\mathbb{P}_n\varphi(\boldsymbol{O}; \widehat{\boldsymbol{\psi}}_{\text{mr}}, \hat{\boldsymbol{\eta}})\varphi(\boldsymbol{O}; \widehat{\boldsymbol{\psi}}_{\text{mr}}, \widehat{\boldsymbol{\eta}})^T\}(\widehat{\boldsymbol{M}}^{-1})^T$, with $\widehat{\boldsymbol{M}} = \mathbb{P}_n \partial\widehat{\varphi}(\boldsymbol{O}; \boldsymbol{\psi}, \widehat{\boldsymbol{\eta}}) / \partial\boldsymbol{\psi}|_{\psi = \widehat{\psi}_{\text{mr}}}$. Even if Equation (5) does not hold, for example, when only one of $(\mathscr{M}_1, \mathscr{M}_2, \mathscr{M}_3)$ holds, but all the nuisance parameters are finite-dimensional and in the form of M-estimators, $\widehat{\boldsymbol{\psi}}_{\text{mr}}$ is still consistent and asymptotically normal from standard M-estimation theory (Newey & McFadden, 1994, Chap. 6). Thus, a consistent variance estimator for $\sqrt{n}\widehat{\boldsymbol{\psi}}_{\text{mr}}$ can be constructed by stacking the efficient influence function $\varphi(\boldsymbol{O}; \boldsymbol{\psi}, \boldsymbol{\eta})$ and the estimation equations for the nuisance parameters, solving for $(\widehat{\boldsymbol{\psi}}_{\text{mr}}, \widehat{\boldsymbol{\eta}})$ simultaneously, and taking the corresponding diagonal component of the joint sandwich variance estimator. Alternatively, the nonparametric bootstrap is commonly used in practice (Cheng & Huang, 2010).

4 | TWO-SAMPLE INSTRUMENTED DIFFERENCE-IN-DIFFERENCES

In some applications, it is hard to collect the exposure and outcome variables for the same individual, especially when the outcome is defined to reflect a delayed treatment effect. For instance, in the smoking and lung cancer example in Section 1, the outcome of interest is lung cancer mortality after 35 years and it is infeasible to follow the same individuals for 35 years. Motivated from Angrist and Krueger (1992, 1995)'s influential two-sample standard IV analysis, we extend the instrumented DID to a two-sample design.

Suppose there are n_a i.i.d. realizations of (T_a, Z_a, D_a, Y_a) from one sample, and n_b i.i.d. realizations of (T_b, Z_b, D_b, Y_b) from another sample. These two samples are independent of each other and we never observe D_a and Y_b. We write the observed data as $(T_{ai}, Z_{ai}, Y_{ai}, i = 1, \ldots, n_a)$ and $(T_{bi}, Z_{bi}, D_{bi}, i = 1, \ldots, n_b)$, which are respectively referred to as the outcome dataset and the exposure dataset. Let δ_{Ya}, $\widehat{\delta}_{Ya}$, δ_{Db}, $\widehat{\delta}_{Db}$, $\widehat{\mu}_{Ya}(t, z)$, $\widehat{\mu}_{Db}(t, z)$ be as defined in Equations (1) and (2) but evaluated correspondingly using the outcome dataset and exposure dataset. Suppose that Assumptions 1 and 2 hold for the data-generating processes in both datasets, and $E(Y_a \mid T_a, Z_a) = E(Y_b \mid T_b, Z_b)$, $E(D_a \mid T_a, Z_a) = E(D_b \mid T_b, Z_b)$, then the ATE is identified by $\beta_0 = \delta_{Ya}/\delta_{Db}$. Analogously, the two-sample instrumented DID Wald estimator is obtained as $\widehat{\beta}_{\text{TSwald}} = \widehat{\delta}_{Ya}/\widehat{\delta}_{Db}$. In Theorem S2, we establish the consistency and asymptotic normality of $\widehat{\beta}_{\text{TSwald}}$ and provide a consistent variance estimator. Both $\widehat{\beta}_{\text{TSwald}}$ and its variance estimator can be conveniently calculated based on solely summary statistics $\widehat{\mu}_{Ya}(t, z)$ and $\widehat{\mu}_{Db}(t, z)$ and their standard errors (SEs).

5 | MEASURE OF WEAK IDENTIFICATION

Weak identification is a general challenge for IV-type methods and has recently received increased attention among theoretical and applied researchers; see Stock et al. (2002) for a survey. For standard IV, weak identification refers to that the IVs are only weakly associated with the exposure. For instrumented DID, weak identification refers to that the trends in exposure for $Z = 0$ and $Z = 1$ are near-parallel. Under weak identification, the sampling distribution for the point estimators is generally non-normal and the standard inference can be unreliable (Bound et al., 1995). Therefore, it is important to have a measure of weak identification tailored for the instrumented DID as diagnostic checks to make sure the developed asymptotic inference procedures can be reliably applied.

Consider first the case when there are no observed covariates. We take the one-sample estimator $\widehat{\beta}_{\text{wald}}$ as an example; the result for the two-sample estimator $\widehat{\beta}_{\text{TSwald}}$ is similar. Note that $\widehat{\delta}_Y$ and $\widehat{\delta}_D$ can be respectively obtained from fitting a saturated model of Y or D on

1, ZT, Z, and T, where ZT is the interaction term. Let $\boldsymbol{R}$ be the n-dimensional vector of residuals from regressing ZT on 1, Z, and T. By using the Frisch–Waugh–Lovell theorem (Davidson & MacKinnon, 1993), $\widehat{\beta}_{\text{wald}}$ in Equation (2) can be equivalently formulated as

$$\widehat{\beta}_{\text{wald}} = \frac{\widehat{\delta}_Y}{\widehat{\delta}_D} = \frac{\left(\boldsymbol{R}^T\boldsymbol{R}\right)^{-1}\boldsymbol{R}^T\boldsymbol{Y}}{\left(\boldsymbol{R}^T\boldsymbol{R}\right)^{-1}\boldsymbol{R}^T\boldsymbol{D}} = \frac{\boldsymbol{D}^T\boldsymbol{H}_R\boldsymbol{Y}}{\boldsymbol{D}^T\boldsymbol{H}_R\boldsymbol{D}},$$

where $\boldsymbol{D}^T = (D_1, \ldots, D_n)$, $\boldsymbol{Y}^T = (Y_1, \ldots, Y_n)$, $\boldsymbol{H}_R = \boldsymbol{R}\left(\boldsymbol{R}^T\boldsymbol{R}\right)^{-1}\boldsymbol{R}^T$ is the hat matrix. Interestingly, the above formula indicates that $\widehat{\beta}_{\text{wald}}$ can be alternatively obtained from a conventional two-stage least squares: the exposure $\boldsymbol{D}$ is first regressed on $\boldsymbol{R}$ (first-stage regression) and the outcome $\boldsymbol{Y}$ is then regressed on the predicted values from the first-stage regression. This provides a perception that Z as an IV for DID is equivalent to using ZT as the standard IV while further controlling for 1, Z, and T. Hence, the concentration parameter of ZT as the standard IV (controlling for 1, Z, and T) serves here as a measure of weak identification using Z as the IV for DID. Specifically, this measure is defined as $\kappa^2 = \delta_D^2\boldsymbol{R}^T\boldsymbol{R}/\sigma_\epsilon^2$, where δ_D is defined in Proposition 1, σ_ϵ^2 is the population residual variance from the first-stage regression. Heuristically, κ^2 increases if we have a larger sample size n, larger δ_D^2, or a larger limit of $\boldsymbol{R}^T\boldsymbol{R}/n$. For the usual inference based on normal approximation to be accurate, κ^2 must be large.

A commonly used estimate of κ^2 is the F-statistic from the first-stage regression. When only summary-data are available, that is, only $\widehat{\delta}_D$ and its SE are available, one can also use the squared z-score as an estimate of κ^2, where the z-score is the ratio of $\widehat{\delta}_D$ to its SE. When there are observed covariates, a measure of weak identification can also be easily calculated by defining $\boldsymbol{R}$ as the vector of residuals from regressing ZT on 1, Z, T, $\boldsymbol{X}$. We follow Stock et al. (2002) and recommend checking to make sure that an estimated κ^2 is larger than 10 before applying the inference methods in Sections 3 and 4.

6 | SIMULATIONS

To evaluate the finite sample performance of the proposed instrumented DID (iDID) methods, we simulate data as follows: $\boldsymbol{X} = (X_1, X_2)^T$, $X_1 \sim N(0,1)$, $X_2 \sim N(0,1)$, $Z \sim \text{Binom}(\text{expit}(0.5I_{X_1 > 0} + 0.5I_{X_2 > 0}))$, $T \sim \text{Binom}(0.5)$, $U_t \sim N(2t - 1,1)$, $\epsilon_t \sim N(0,1)$, $D_t \sim \text{Binom}(\text{expit}(-0.5 - ZU_t + 1.5U_t))$, $Y_t = (1 + X_1 + X_2)D_t + 2 + 2U_t + Z + (1 + X_1 + X_2) + \epsilon_t$, for $t = 0,\ 1$. We simulate $n = 10^5$ random samples from $(T, Z, \boldsymbol{X}, D_0, D_1, Y_0, Y_1)$ and let $D = TD_1 + (1 - T)D_0$, $Y = TY_1 + (1 - T)Y_0$. The observed data are $(Z_i, \boldsymbol{X}_i, T_i, D_i, Y_i, i = 1, \ldots, n)$.

Under this data-generating process, Assumptions 1 and 2 do not hold unconditionally, but do hold in each of the four strata defined by $(I_{X_1 > 0}, I_{X_2 > 0})$. Hence, the Wald estimator in Equation (2) is valid when considering each stratum separately; we denote the obtained stratum-specific Wald estimators as $\widehat{\beta}_{\text{S1,wald}}, \ldots, \widehat{\beta}_{\text{S4,wald}}$ and they are respectively estimating the stratum-specific ATE: −0.60, 1, 1, 2.60. On the other hand, Assumptions 1 and 2 hold when conditioning on $\boldsymbol{X}$, and thus the three classes of semiparametric estimators $\widehat{\boldsymbol{\psi}}_{\mathrm{reg}}$, $\widehat{\boldsymbol{\psi}}_{\mathrm{ipw}}$, $\widehat{\boldsymbol{\psi}}_{\mathrm{g}}$ and the multiply robust estimator $\widehat{\boldsymbol{\psi}}_{\mathrm{mr}}$ proposed in Section 3.2 are all valid. For the semiparametric iDID method, we consider two *working models* for the CATE: a constant working model $\beta(\boldsymbol{v}; \boldsymbol{\psi}) = \psi$ with $\boldsymbol{V} = 1$ and a linear working model $\beta(\boldsymbol{v}; \boldsymbol{\psi}) = \psi_1 + \psi_2 x_1$, with $\boldsymbol{V} = X_1$. The true values of $\psi,\ \psi_1,\ \psi_2$ are all equal to 1 because $E(Y_t^{(1)} - Y_t^{(0)}) = 1$ and $(Y_t^{(1)} - Y_t^{(0)} \mid X_1) = 1 + X_1$. The weight function $w(\boldsymbol{v})$ in Equation (3) is set to be 1.

We also examine the effect of model misspecification for the semiparametric iDID estimators. Note that the data-generating process implies that $(t, 1, \boldsymbol{x}) = \mathrm{expit}(0.5\boldsymbol{I}_{x_1 > 0} + 0.5\boldsymbol{I}_{x_2 > 0})/2,\ \ \pi(t, 0, \boldsymbol{x}) = \{1 - \mathrm{expit}(0.5\boldsymbol{I}_{x_1 > 0} + 0.5\boldsymbol{I}_{x_2 > 0})\}/2$, and $\delta_D(\boldsymbol{x})$, $\delta(\boldsymbol{x})$, $b_D(\boldsymbol{x})$, $b_Y(\boldsymbol{x})$, $m_{DZ}(\boldsymbol{x})$, $m_{DT}(\boldsymbol{x})$, $m_{YZ}(\boldsymbol{x})$, $m_{YT}(\boldsymbol{x})$ are all linear in $\boldsymbol{x}$. The misspecified model we fit for $\pi(t, z, \boldsymbol{x})$ is a product of two logistic regressions, one for Z and the other for T, both in terms of $\exp(x_1/2)$, the misspecified models for $\delta_D(\boldsymbol{x})$, $\delta(\boldsymbol{x})$ are linear in x_1, and for $b_D(\boldsymbol{x})$, $b_Y(\boldsymbol{x})$, $m_{DZ}(\boldsymbol{x})$, $m_{DT}(\boldsymbol{x})$, $m_{YZ}(\boldsymbol{x})$, $m_{YT}(\boldsymbol{x})$ are linear in $\exp(x_1/2)$.

We compare with two other methods, direct treated-versus-control outcome comparison using ordinary least squares (OLS) and the standard IV method using Z as the IV, where the latter is implemented using the R package ivpack (Jiang & Small, 2014). Direct outcome comparison is invalid because of the unmeasured confounder U_t; the standard IV method is also invalid due to the direct effect of Z on the outcome. Table 1 shows the simulation results based on 1,000 repetitions, which includes: (i) the simulation average bias and standard deviation (SD) of each estimator; (ii) the mean standard errors (SEs), which are calculated according to Equation (S4) in the supplementary materials for the Wald estimator, using standard M-estimation theory for the semiparametric estimators; (iii) simulation coverage probability (CP) of 95% confidence intervals.

The following is a summary based on the results in Table 1. The OLS and standard IV estimators have large bias due to violations of their assumptions. The stratum-specific iDID Wald estimators show negligible bias and adequate coverage probability. The three classes of semiparametric iDID estimators that rely on $\mathscr{M}_1,\ \mathscr{M}_2,\ \mathscr{M}_3$ have negligible bias and adequate coverage probability when the corresponding models are correctly specified but are biased when misspecified. The multiply robust semiparametric iDID estimators exhibit negligible bias and adequate coverage probabilities when at least one of $(\mathscr{M}_1, \mathscr{M}_2, \mathscr{M}_3)$ is correct, which supports the multiple robustness property.

7 | APPLICATION

We apply the proposed method to analyze the effect of cigarette smoking on lung cancer mortality. Given the lag between smoking exposure and lung cancer mortality, we adopt the two-sample instrumented DID design. Our analysis is based upon two datasets arranged by 10-year birth cohort: the 1970 National Health Interview Survey (NHIS) for nationally representative estimates of smoking prevalence (National Health Interview Survey, 1970), and the US Centers for Disease Control and Prevention’s (CDC) Wide-ranging ONline Data for Epidemiologic Research (WONDER) system for estimates of national lung cancer (ICD-8/9: 162; ICD-10: C33-C34) mortality rates (CDC, 2000a, 2000b, 2016). Only the

1970 NHIS is used because it is the first NHIS that records the initiation and cessation time of smoking such that a longitudinal structure is available. We closely follow the approach taken by Tolley et al. (1991, Chapter 3) to calculate the smoking prevalence rates.

Based on the data availability, we focus on four successive 10-year birth cohorts: 1911–1920, 1921–1930, 1931–1940, 1941–1950, whose smoking prevalence is estimated respectively at year 1940, 1950, 1960, 1970 when they are at age 20–29, whose lung cancer mortality rates are estimated respectively at year 1975, 1985, 1995, 2005 when they are at age 55–64. Here, cohort of birth plays the role of time. Figure 2 shows the changes in prevalence of cigarette smoking among men and women aged 20–29 years, and the changes in lung cancer mortality rates 35 years later in the United States. From Figure 2, we see that the trends in lung cancer mortality rates follow the trends in smoking prevalence, with a lag of 35 years, which provides evidence that smoking increases lung cancer mortality rate.

There have been many direct comparisons of the lung cancer mortality rates between smokers and non-smokers which have found higher rates among smokers (International Agency for Research on Cancer, 1986). Additional studies that replicate direct comparisons of smokers and non-smokers may not add much evidence beyond the first comparison. It is argued in Rosenbaum (2010) that “in such a situation, it may be possible to find haphazard nudges that, at the margin, enable or discourage [the exposure]. ... These nudges may be biased in various ways, but there may be no reason for them to be consistently biased in the same direction, so similar estimates of effect from studies subject to different potential biases gradually reduce ambiguity about what part is effect and what part is bias.” The instrumented DID is one such method that attempts to exploit the “haphazard nudges”, that is, the targeted tobacco advertising to women in the 1960s that led to a rapid increase in smoking among young women in a way that is presumably independent of other causes of lung cancer mortality.

To quantitatively evaluate the effect of cigarette smoking on lung cancer mortality, we take gender—a surrogate of whether each individual received encouragement (targeted tobacco advertising) or not—as the IV for DID. Note that gender does not need to have a causal effect on smoking; as proved in the Supporting information, it suffices that gender is correlated with smoking due to the encouragement from targeted tobacco advertising. We consider two successive 10-year birth cohorts, setting the earlier birth cohort as $T = 0$ and the later birth cohort as $T = 1$. Gender is likely a valid IV for DID, as it clearly satisfies the trend relevance assumption, the lung cancer mortality rates for men and women would have evolved similarly had all subjects counterfactually not smoked, and there is no evident gender difference in the cancer-causing effects of cigarette smoking (Patel et al., 2004).

Table 2 summarizes (i) the F-statistic proposed in Section 5 to measure weak identification; and (ii) the two-sample iDID Wald estimators $\widehat{\beta}_{\mathrm{TSwald}}$ defined in Section 4 and their SEs defined in Equation (S6). More details on the application are also in the Supporting information. From Table 2, under the assumption that gender is a valid IV for DID and the treatment effect is stable over time, we find evidence that smoking leads to significantly higher lung cancer mortality rates. Specifically, we find that smoking in one’s 20s leads to an elevated annual lung cancer mortality rate at age 55–64 years, with the effect size ranging from 0.285% to 0.568%. This is of a similar magnitude as the findings in Thun et al. (1982, 2013). Using different birth cohorts gives slightly different point estimates, but they are within two SEs of each other. Nonetheless, there is still concern about violating the stable treatment effect over time assumption (Assumption 2(d)), possibly because the cigarette design and composition have undergone changes that promote deeper inhalation of smoke (Thun et al., 2013; Warren et al., 2014). In Section S4, we perform a sensitivity analysis and find that increasing risk of smoking over time does not explain away the observed treatment effect.

8 | RESULTS AND DISCUSSION

In this paper, we have proposed a new method called instrumented DID that explicitly leverages exogenous randomness in the exposure trends, and controls for unmeasured confounding in repeated cross-sectional studies. The instrumented DID method evolves from two powerful natural experiment devices, the standard IV and standard DID, but is able to relax some of their most disputable assumptions. Our motivation of assessing the causal effect by linking the change in outcome mean and the change in exposure rate is also related to the trend-in-trend design (Ji et al., 2017) and etiologic mixed design (Lash et al., 2021).

In principle, any variable that satisfies Assumptions 2(a)–(c) can be chosen as the IV for DID. Here, we list two common sources of the IV for DID: (i) administrative information, such as geographic region and insurance type; and (ii) variables that are commonly used as standard IVs, such as physician preference, distance to care provider, and genetic variants—see Baiocchi et al. (2014) for more examples; as discussed in Section 2, these variables are more likely to be an IV for DID compared to being a standard IV, because IVs for DID are allowed to have direct effects on the outcome.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGMENTS

We gratefully acknowledge support by grant R01AG064589 from the National Institutes of Health. We would like to thank the anonymous referees, an Associate Editor and the Editor for their constructive comments that led to a much improved paper.

Funding information

National Institutes of Health, Grant/Award Number: R01AG064589

DATA AVAILABILITY STATEMENT

The data that support the findings of this paper are openly available in a GitHub repository at https://github.com/jfiksel/compregpaper.

REFERENCES

Abadie A (2003) Semiparametric instrumental variable estimation of treatment response models. Journal of Econometrics, 113, 231–263.

Abadie A (2005) Semiparametric difference-in-differences estimators. The Review of Economic Studies, 72, 1–19.

Angrist JD, Imbens GW & Rubin DB (1996) Identification of causal effects using instrumental variables. Journal of the American Statistical Association, 91, 444–455.

Angrist JD & Krueger AB (1992) The effect of age at school entry on educational attainment: an application of instrumental variables with moments from two samples. Journal of the American statistical Association, 87, 328–336.

Angrist JD & Krueger AB (1995) Split-sample instrumental variables estimates of the return to schooling. Journal of Business & Economic Statistics, 13, 225–235.

Angrist JD & Pischke J-S (2008) Mostly harmless econometrics: an empiricist's companion. Princeton, NJ: Princeton University Press.

Bailar JC & Gornik HL (1997) Cancer undefeated. New England Journal of Medicine, 336, 1569–1574. [PubMed: 9164814]

Baiocchi M, Cheng J & Small DS (2014) Instrumental variable methods for causal inference. Statistics in Medicine, 33, 2297–2340. [PubMed: 24599889]

Bickel P, Klaassen C, Ritov Y & Wellner J (1993) Efficient and adaptive estimation for semiparametric Models. Springer.

Burbank F (1972) U.S. lung cancer death rates begin to rise proportionately more rapidly for females than for males: a dose-response effect? Journal of Chronic Diseases, 25, 473–479. [PubMed: 4636342]

Card D & Krueger AB (1994) Minimum wages and employment: a case study of the fast food industry in new jersey and pennsylvania. American Economic Review, 84, 772–793.

CDC (2000a). Centers for disease control and prevention, national center for health statistics. compressed mortality file 1968–1978. CDC WONDER online database, compiled from compressed mortality file CMF 1968–1988, series 20, no. 2A, 2000. Available from: http://wonder.cdc.gov/cmf-icd8.html [Accessed 27th Aug 2020].

CDC (2000b). Centers for disease control and prevention, national center for health statistics. compressed mortality file 1979–1998. CDC WONDER online database, compiled from compressed mortality file CMF 1979–1998, series 20, no. 2A, 2000 and CMF 1989–1998, series 20, no. 2E, 2003. Available from: http://wonder.cdc.gov/cmf-icd9.html [Accessed 27th Aug 2020].

CDC (2016) Centers for disease control and prevention, national center for health statistics. compressed mortality file 1999–2016 on cdc wonder online database, released june 2017. Data are from the compressed mortality file 1999–2016 series 20 no. 2U, 2016. Available from: http://wonder.cdc.gov/cmf-icd10.html [Accessed 28th Aug 2020].

Cheng G & Huang JZ (2010) Bootstrap consistency for general semiparametric M-estimation. The Annals of Statistics, 38, 2884–2915.

Cui Y & Tchetgen Tchetgen E (2021) A semiparametric instrumental variable approach to optimal treatment regimes under endogeneity. Journal of the American Statistical Association, 116, 162–173. [PubMed: 33994604]

Davidson R & MacKinnon JG (1993) Estimation and inference in econometrics. Oxford University Press.

de Chaisemartin C & D'HaultfŒuille X (2018) Fuzzy differences-in-differences. The Review of Economic Studies, 85, 999–1028.

Duflo E (2001) Schooling and labor market consequences of school construction in indonesia: evidence from an unusual policy experiment. American Economic Review, 91, 795–813.

Hernán MA & Robins JM (2006) Instruments for causal inference: an epidemiologist's dream? Epidemiology, 17, 360–372. [PubMed: 16755261]

Hernan MA & Robins JM (2020) Causal inference: what if. Boca Raton, FL: Chapman & Hall.

International Agency for Research on Cancer (1986) Tobacco smoking, vol. 38. World Health Organization.

Ji X, Small DS, Leonard CE & Hennessy S (2017) The trend-in-trend research design for causal inference. Epidemiology, 28, 529–536. [PubMed: 27775954]

Jiang Y & Small DS (2014) ivpack: instrumental Variable Estimation. R package version 1.2.

Kennedy EH, Lorch S & Small DS (2019) Robust causal inference with continuous instruments using the local instrumental variable curve. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 81, 121–143.

Lash TL, VanderWeele TJ, Haneuse S & Rothman KJ (2021) Modern epidemiology, vol. 4. Wolters Kluwer Health.

Lawlor DA, Davey Smith G, Kundu D, Bruckdorfer KR & Ebrahim S (2004) Those confounded vitamins: what can we learn from the differences between observational versus randomised trial evidence? Lancet, 363, 1724–1727. [PubMed: 15158637]

Meigs JW (1977) Epidemic lung cancer in women. JAMA, 238, 1055–1055. [PubMed: 577950]

National Health Interview Survey (1970) Available from: ftp://ftp.cdc.gov/pub/health_statistics/nchs/datasets/nhis/1970 [Accessed 31st Aug 2020].

Newey WK & McFadden D (1994) Large sample estimation and hypothesis testing. Chap. 36. Handbook of Econometrics, 4, 2111–2245.

Neyman J (1923) On the application of probability theory to agricultural experiments. Essay on principles. section 9. Statistical Science, 5, 465–472. Trans. Dabrowska Dorota M. and Speed Terence P. (1990).

Ogburn EL, Rotnitzky A & Robins JM (2015) Doubly robust estimation of the local average treatment effect curve. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 77, 373–396. [PubMed: 25663814]

Patel JD, Bach PB & Kris MG (2004) Lung cancer in US women: a contemporary epidemic. JAMA, 291, 1763–1768. [PubMed: 15082704]

Pierce JP & Gilpin EA (1995) A historical analysis of tobacco marketing and the uptake of smoking by youth in the United States: 1890–1977. Health Psychology, 14, 500. [PubMed: 8565924]

Robins JM (1994) Correcting for non-compliance in randomized trials using structural nested mean models. Communications in Statistics: Theory and Methods, 23, 2379–2412.

Rosenbaum PR (2010) Design of observational studies. Springer.

Rubin DB (1974) Estimating causal effects of treatments in randomized and nonrandomized studies. Journal of Educational Psychology, 6, 688–701.

Rubin DB (1978) Bayesian inference for causal effects: the role of randomization. Annals of Statistics, 6, 34–58.

Rubin DB (1990) Comment: Neyman (1923) and causal inference in experiments and observational studies. Statistical Science, 5, 472–480.

Rutter M (2007) Identifying the environmental causes of disease: how should we decide what to believe and when to take action? Report Synopsis. Academy of Medical Sciences.

Shi X, Miao W, Nelson JC & Tchetgen Tchetgen EJ (2020) Multiply robust causal inference with double-negative control adjustment for categorical unmeasured confounding. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 82, 521–540. [PubMed: 33376449]

Stock JH, Wright JH & Yogo M (2002) A survey of weak instruments and weak identification in generalized method of moments. Journal of Business & Economic Statistics, 20, 518–529.

Tan Z (2006) Regression and weighting methods for causal inference using instrumental variables. Journal of the American Statistical Association, 101, 1607–1618.

Thun JM, Day-Lally C, Myers GD, Calle EE, Flanders WD, Zhu B-P et al. (1982) Trends in tobacco smoking and mortality from cigarette use in cancer prevention studies I (1959–1965) and II (1982–1988). Changes in cigarette-related disease risks and their implication for prevention and control: smoking and tobacco control monograph. Vol. 8. Bethesda, MD: U.S. Department of Health and Human Services, National Institutes of Health, National Cancer Institute. NIH Pub.

Thun MJ, Carter BD, Feskanich D, Freedman ND, Prentice R, Lopez AD, et al. (2013) 50-year trends in smoking-related mortality in the United States. New England Journal of Medicine, 368, 351–364. [PubMed: 23343064]

Tolley H, Crane L & Shipley N (1991) Strategies to control tobacco use in the United States—a blueprint for public health action in the 1990s. NIH publication no. 92–3316 pp. 75–144. Bethesda, MD: U.S. Department of Health and Human Services, Public Health Service, National Institutes of Health, National Cancer Institute.

van der Vaart A (2000) Asymptotic statistics. Cambridge University Press.

Vansteelandt S, VanderWeele TJ, Tchetgen Tchetgen EJ & Robins JM (2008) Multiply robust inference for statistical interactions. Journal of the American Statistical Association, 103, 1693–1704. [PubMed: 21603124]

Wang L & Tchetgen Tchetgen E (2018) Bounded, efficient and multiply robust estimation of average treatment effects using instrumental variables. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 80, 531–550. [PubMed: 30034269]

Warren GW, Alberg AJ, Kraft AS & Cummings KM (2014) The 2014 surgeon general's report:"the health consequences of smoking—50 years of progress": a paradigm shift in cancer care. Cancer, 120, 1914–1916. [PubMed: 24687615]

Wooldridge JM (2010) Econometric analysis of cross section and panel data. MIT press.

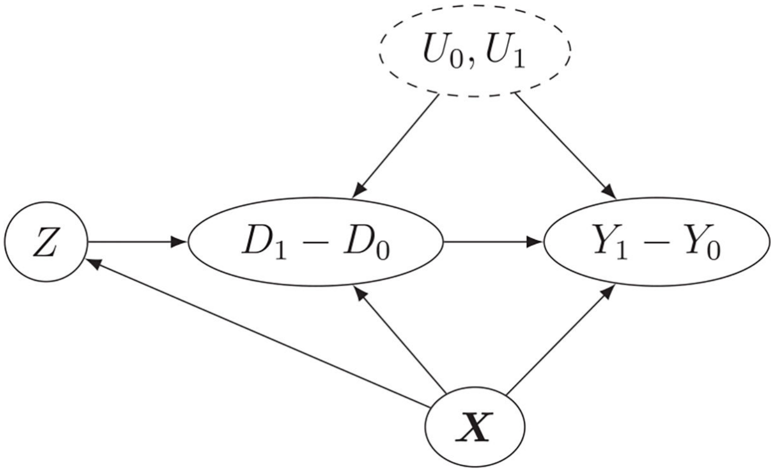

FIGURE 1.

Directed acyclic graph (DAG) for instrumented difference-in-differences (DID). Suppose the existence of an unmeasured confounder U_t such that $(D_0, D_1) \perp (Y_0, Y_1) | U_0,\ U_1,\ \boldsymbol{X}$. Assumption 2(a) states that Z must be associated with the change in exposure $D_1 - D_0$, Assumption 2(b) states that Z is independent of any unmeasured confounders $U_0,\ U_1$ and cannot have any direct effect on the change in outcome $Y_1 - Y_0$ and does not modify the treatment effect.

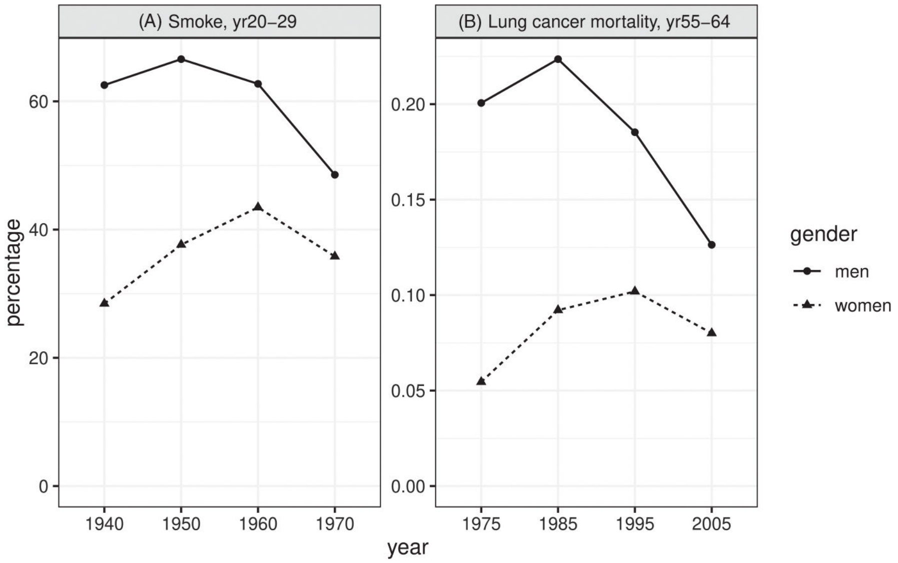

FIGURE 2.

Changes in prevalence of cigarette smoking for men and women aged 20–29, lung cancer mortality rates for men and women aged 55–64 years among four successive 10-year birth cohorts: 1911–1920, 1921–1930, 1931–1940, 1941–1950

TABLE 1

Bias, standard deviation (SD), average standard error (SE), and coverage probability (CP) of 95% asymptotic confidence interval based on 1,000 repetitions with $n = 10^5$

Method	Correct model	Estimator	Bias	SD	SE	CP
OLS			2.466	0.024	0.023	0.000
Standard IV			−38.525	3.385	3.316	0.000
iDID-Wald		$\widehat{\beta}_{\mathrm{S1,\,wald}}$	−0.014	0.247	0.251	0.950
		$\widehat{\beta}_{\mathrm{S2,\,wald}}$	0.007	0.253	0.259	0.958
		$\widehat{\beta}_{\mathrm{S3,\,wald}}$	−0.008	0.250	0.259	0.961
		$\widehat{\beta}_{\mathrm{S4,\,wald}}$	0.000	0.289	0.284	0.943
Constant working model$\beta(\boldsymbol{V};\boldsymbol{\psi}) = \psi$						
iDID-MR	$(\mathscr{M}_1, \mathscr{M}_2, \mathscr{M}_3)$ all correct	$\widehat{\psi}_{\mathrm{mr}}$	−0.002	0.111	0.114	0.956
	$\mathscr{M}_1$ correct	$\widehat{\psi}_{\mathrm{mr}}$	−0.001	0.110	0.114	0.960
	$\mathscr{M}_2$ correct	$\widehat{\psi}_{\mathrm{mr}}$	−0.003	0.136	0.139	0.944
	$\mathscr{M}_3$ correct	$\widehat{\psi}_{\mathrm{mr}}$	−0.003	0.137	0.140	0.945
	none	$\widehat{\psi}_{\mathrm{mr}}$	−0.355	0.144	0.142	0.293
iDID-Reg	$\mathscr{M}_1$ correct	$\widehat{\psi}_{\mathrm{reg}}$	−0.002	0.109	0.114	0.960
	$\mathscr{M}_1$ incorrect	$\widehat{\psi}_{\mathrm{reg}}$	−0.351	0.144	0.149	0.335
iDID-IPW	$\mathscr{M}_2$ correct	$\widehat{\psi}_{\mathrm{ipw}}$	−0.021	0.225	0.225	0.948
	$\mathscr{M}_2$ incorrect	$\widehat{\psi}_{\mathrm{ipw}}$	−0.271	0.234	0.242	0.816
iDID-G	$\mathscr{M}_3$ correct	$\widehat{\psi}_{\mathrm{g}}$	−0.021	0.225	0.224	0.948
	$\mathscr{M}_3$ incorrect	$\widehat{\psi}_{\mathrm{g}}$	−0.276	0.235	0.233	0.814
Linear working model$\beta(\boldsymbol{V};\boldsymbol{\psi}) = \psi_1 + \psi_2 X_1$						
iDID-MR	$(\mathscr{M}_1, \mathscr{M}_2, \mathscr{M}_3)$ all correct	$\widehat{\psi}_{1,\,\mathrm{mr}}$	−0.002	0.110	0.114	0.956
		$\widehat{\psi}_{2,\,\mathrm{mr}}$	0.004	0.113	0.115	0.950
	$\mathscr{M}_1$ correct	$\widehat{\psi}_{1,\,\mathrm{mr}}$	−0.001	0.110	0.114	0.960
		$\widehat{\psi}_{2,\,\mathrm{mr}}$	0.004	0.115	0.118	0.946
	$\mathscr{M}_2$ correct	$\widehat{\psi}_{1,\,\mathrm{mr}}$	−0.003	0.136	0.139	0.944

Method	Correct model	Estimator	Bias	SD	SE	CP
		$\widehat{\psi}_{2,\mathrm{mr}}$	0.003	0.146	0.150	0.960
	$\mathscr{M}_3$ correct	$\widehat{\psi}_{1,\mathrm{mr}}$	−0.003	0.137	0.140	0.946
		$\widehat{\psi}_{2,\mathrm{mr}}$	0.004	0.144	0.149	0.958
	None	$\hat{\psi}_{1,\mathrm{mr}}$	−0.355	0.144	0.142	0.292
		$\widehat{\psi}_{2,\mathrm{mr}}$	−0.129	0.221	0.175	0.908
iDID-Reg	$\mathscr{M}_1$ correct	$\widehat{\psi}_{1,\mathrm{reg}}$	−0.001	0.109	0.114	0.960
		$\widehat{\psi}_{2,\mathrm{reg}}$	−0.005	0.114	0.118	0.949
	$\mathscr{M}_1$ incorrect	$\widehat{\psi}_{1,\mathrm{reg}}$	−0.351	0.144	0.149	0.332
		$\widehat{\psi}_{2,\mathrm{reg}}$	−0.110	0.473	0.478	0.942
iDID-IPW	$\mathscr{M}_2$ correct	$\widehat{\psi}_{1,\mathrm{ipw}}$	−0.021	0.225	0.225	0.948
		$\widehat{\psi}_{2,\mathrm{ipw}}$	−0.010	0.270	0.269	0.957
	$\mathscr{M}_2$ incorrect	$\widehat{\psi}_{1,\mathrm{ipw}}$	−0.271	0.234	0.242	0.813
		$\widehat{\psi}_{2,\mathrm{ipw}}$	−0.072	0.267	0.313	0.957
iDID-G	$\mathscr{M}_3$ correct	$\widehat{\psi}_{1,\mathrm{g}}$	−0.021	0.225	0.224	0.949
		$\widehat{\psi}_{2,\mathrm{g}}$	−0.007	0.245	0.247	0.953
	$\mathscr{M}_3$ incorrect	$\widehat{\psi}_{1,\mathrm{g}}$	−0.276	0.235	0.233	0.812
		$\widehat{\psi}_{2,\mathrm{g}}$	−0.234	0.246	0.241	0.863

Abbreviations: iDID, instrumented difference-in-differences; OLS, ordinary least squares.

TABLE 2

Two-sample iDID Wald estimates and their standard errors (in parentheses) using two successive birth cohorts (in %)

	1911–1920	**1921–1930**	**1931–1940**
Birth cohort	**1921–1930**	**1931–1940**	**1941–1950**
F-statistic	13.94	47.28	21.33
$\widehat{\beta}_{\mathrm{TSwald}}$	0.285 (0.089)	0.497 (0.076)	0.568 (0.127)

0*Notes:* F-statistic is the squared z-score, $\widehat{\beta}_{\mathrm{TSwald}}$ defined in Section 4 estimates the ATE of smoking on lung cancer mortality. Abbreviations: ATE, average treatment effect; iDID, instrumented difference-in-differences.