Electronic Journal of Statistics
Vol. 18 (2024) 5132–5155
ISSN: 1935-7524
https://doi.org/10.1214/24-EJS2313

Structural mean models for instrumented difference-in-differences

Tat-Thang Vo

Research group EPIDERME
Faculty of Medicine, University Paris Est Creteil
e-mail: tat-thang.vo@u-pec.fr

Ting Ye

Department of Biostatistics
Hans Rosling Center for Population Health, University of Washington
e-mail: tingye1@uw.edu

Ashkan Ertefaie

Department of Biostatistics and Computational Biology
University of Rochester
e-mail: ashkan_ertefaie@urmc.rochester.edu

Samrat Roy

Operations and Decision Sciences
Indian Institute of Management Ahmedabad
e-mail: samratr@iima.ac.in

James Flory

Department of Subspecialty Medicine
Memorial Sloan Kettering Cancer Center
e-mail: floryj@mskcc.org

Sean Hennessy

Department of Biostatistics, Epidemiology and Informatics
Perelman School of Medicine, University of Pennsylvania
e-mail: hennessy@pennmedicine.upenn.edu

Stijn Vansteelandt

Department of Applied Mathematics, Computer Science and Statistics
Ghent University
e-mail: stijn.vansteelandt@ugent.be

and

Dylan S Small

Department of Statistics and Data Science
The Wharton School, University of Pennsylvania,
e-mail: dsmall@wharton.upenn.edu

Abstract: In the standard difference-in-differences research design, the parallel trend assumption can be violated when the effect of some unmeasured confounders on the outcome trend is different between the treated and untreated populations. Progress can be made if there is an exogenous variable that (i) does not directly influence the change in outcome (i.e. the outcome trend) except through influencing the change in exposure (i.e. the exposure trend), and (ii) is not related to the unmeasured exposure - outcome confounders on the trend scale. Such exogenous variable is called an instrument for difference-in-differences. For continuous outcomes that lend themselves to linear modelling, so-called instrumented difference-in-differences methods have been proposed. In this paper, we will suggest novel multiplicative structural mean models for instrumented difference-in-differences, which allow one to identify and estimate the average treatment effect that is stable over time on the multiplicative scale, in the whole population or among the treated, when (i) a valid instrument for difference-in-differences is available and (ii) there is no carry-over effect across periods. We discuss the identifiability of these models, then develop efficient semi-parametric estimation approaches that allow the use of flexible, data-adaptive or machine learning methods to estimate the nuisance parameters. We apply our proposal on health care data to investigate the risk of moderate to severe weight gain under sulfonylurea treatment compared to metformin treatment, among new users of antihyperglycemic drugs.

Keywords and phrases: Difference-in-differences, instrumental variable, semi-parametric theory.

Received January 2023.

Contents

1. Introduction

The estimation of treatment effects in observational studies is often subject to bias due to unmeasured confounding. For instance, observational pharmacoepidemiological studies often utilize data from large administrative claim databases

or electronic health records, which were not collected for research and may have incomplete/inaccurate information on potential confounding variables [39]. In view of this concern, various analytical methods have been proposed to detect or control for unmeasured confounding [28]. Among these approaches, instrumental variable and difference-in-differences designs are very commonly used [2, 9, 37]. Instrumental variable methods make use of an exogeneous variable that is associated with the exposure, but that does not directly affect the outcome and is independent of unmeasured confounders [2]. The difference-in-differences method is instead 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. Assuming that the outcomes in the two exposure groups evolve in the same way over time in the absence of the exposure (i.e., the parallel trends assumption), the difference-in-differences method is able to remove time-invariant bias caused by unmeasured confounders [37].

To further relax assumptions, the *instrumented difference-in-differences* design has recently been proposed, which combines the strength of instrumental variables and difference-in-differences [38]. This method allows one to identify the treatment effects under a weaker set of assumptions than each parent method alone. As an example, in the standard difference-in-differences design, the parallel trends assumption could be violated when the effect of some unmeasured baseline confounders on the outcome trend in the treated population is different from that in the untreated population, even when these unmeasured confounders do not modify the treatment effect. The instrumented difference-in-differences method overcomes these challenges by (i) allowing for patients to be treated at both timepoints, and (ii) leveraging an exogenous variable that does not have any direct causal impact on the outcome trend except via the exposure trend, and is not associated with the unmeasured confounders on the trend scale [38]. Importantly, this so-called instrument for difference-in-differences need not itself be a valid instrumental variable for the considered exposure-outcome association. For instance, it can have a direct causal effect on the outcome that is not mediated through the exposure at each time point.

Thus far, instrumented difference-in-differences has been developed for settings where the exposure and outcome trends are defined on the additive scale [38]. In many clinical applications with count or binary outcomes, the effect and trends on the multiplicative scale might be of more interest to applied researchers. Akin to standard instrumental variables, identifying multiplcative exposure effects by instrumented difference-in-differences requires non-trivial adjustments of the underlying causal assumptions. Moreover, the obtained identification results in these non-linear settings often involve multiple nuisance parameters that are difficult to estimate. In view of this, extending instrumented difference-in-differences to non-linear settings, and developing flexible estimation strategies that allow for the use of data-driven algorithms can greatly enhance the practical applicability of this method.

In this paper, we aim to improve the utility of instrumented difference-in-differences by proposing structural mean models for this design. Structural mean

models were first introduced by Robins [22] and Robins and Tsiatis [23], and then were extended to instrumental variable and other settings by Vansteelandt and Goetghebeur [33], Hernán and Robins [11], Tchetgen Tchetgen, Robins and Rotnitzky [26], among many others. Our contributions to this literature can be summarized as follows:

First, we propose a set of causal assumptions to identify the average exposure effect among the exposed, defined on the additive scale. We achieve this via considering additive structural mean models for instrumented difference-in-differences. The advantage of these models lies in their ability to offer flexibility in modeling non-linear relationships. Moreover, focusing on the group of exposed individuals also allows one to avoid the assumption of no unmeasured effect modification (and extensions thereof), which is needed in previous works to identify the exposure effect in the whole population [38].

Second, we extend instrumented difference-in-differences to settings with a count outcome or a rare binary outcome. We achieve this by proposing multiplicative structural mean models for instrumented difference-in-differences. As in the additive case, under certain causal assumptions, the proposed multiplicative structural mean models allow one to identify and estimate the average treatment effect on the multiplicative scale, in the whole population or among the exposed, when a valid instrument for difference-in-differences is available.

Third, we develop robust and efficient estimation strategies for the parameters indexing the multiplicative structural mean models, using semi-parametric theory. Proposed estimators can achieve $\sqrt{n}$ rate of convergence to the parameters of interest, even when the nuisance functions are estimated at slower rates. This allows for the utilization of flexible, data-adaptive, or machine learning methods. We consider two different settings. In the first setting, the impact of the baseline covariate on the outcome in the structural mean models is characterized by some finite-dimensional parameter vector. In the second setting, it is left unspecified.

2. Additive structural mean models for instrumented difference-in-differences

Assume that a random sample of a target population is followed up over two time points, i.e. $t = 0$ and $t = 1$. For each individual i in the sample, we observe $O_i = (Z_i, X_i, D_{0i}, Y_{0i}, D_{1i}, Y_{1i})$; where D_{ti} and Y_{ti} are the respective exposure and outcome status observed at each time point t $(t = 0, 1)$, X_i is a vector of baseline covariates and $Z_i = 0, 1$ is a binary instrument for difference-in-differences observed at baseline. The observations $(O_1, \ldots, O_n)$ are independent and identically distributed realizations of $O = (Z, X, D_0, Y_0, D_1, Y_1)$. Note that we allow the presence of exposed patients in the first time point (i.e. $P(D_0 = 1) > 0$), which is not permitted in standard difference-in-differences design. In Figure 1a, we describe the relationship between different variables by a causal diagram.

Denote Y_t^d the counterfactual outcome that would be observed at time point t if the exposure D_t were set to d $(d = 0, 1)$. Throughout the rest of the pa-

per, we will suppose that the following consistency and sequential ignorability assumption holds:

Assumption 1. $Y_t^d = Y_t$ *when* $D_t = d$, *for all* $t, d = 0, 1$.

Assumption 2. *There exists* (U_0, U_1) *possibly unmeasured such that (i)* $Y_0^d \perp\!\!\!\perp D_0 \mid U_0, X, Z$ *and* $Y_1^d \perp\!\!\!\perp D_1 \mid U_0, U_1, Z, X, Y_0, D_0$ *for* $d = 0, 1$.

It is worth noting that the setups of instrumented difference-in-differerences and of standard difference-in-differences are not the same. In the latter, no individuals are exposed at the first time point and some become exposed at the second time point (due to the introduction of a policy in this group). In instrumented difference-in-differences, exposed and unexposed patients may present at both timepoints. Our Assumption 2 then states that the exposure-outcome relationship at each time point is confounded by some unmeasured confounders (U_0 and U_1), such that when the information on these variables was available, one would entirely remove confounding bias. Such an assumption is quite standard in the causal inference literature, and is often made in the analysis of longitudinal or repeatedly measured data [13, 10].

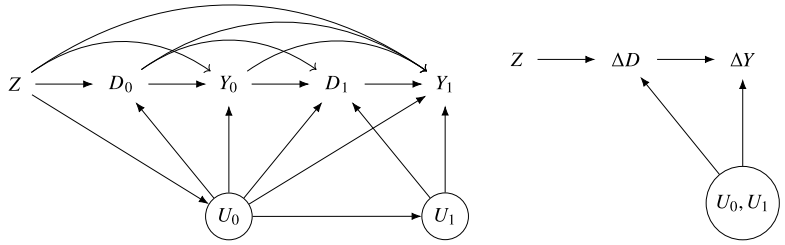


FIG 1. *Data generating mechanism. The baseline covariates X are omitted to simplify the figure. Of note, some causal assumptions needed for identification are not illustrated on these causal diagrams. Readers are referred to the main text for a more detailed discussion of all identification assumptions.*

Our first aim is to make inferences about the conditional average exposure effect given X on the additive scale, assuming that this effect is unchanged over time, and that there is no carry-over effect across timepoints. In clinical medicine and epidemiology, such an assumption is arguably reasonable, especially for treatments of acute conditions, and when the study spans over a short period of time. For example, [17] considered data from a longitudinal study about the incidence of human immunodeficiency virus infection in intravenous drug users from Milan and other areas of northern Italy between 1987 and 1991. As depicted by their Figure 3, they found constant rate differences for HIV infection for parenteral and sexual exposure. Denote $\beta(x)$ the conditional average exposure effect given $X = x$, one then has:

$$\beta(x) = E(Y_t^1 - Y_t^0 | X = x)$$

for $t = 0, 1$. When X is empty or when X does not include any effect modifier, $\beta(x) = \beta$ expresses the average exposure effect.

Under causal diagram 1a, Z cannot be used as a standard instrumental variable to estimate $\beta(x)$. For instance, the exclusion restriction assumption is violated because Z may have a direct effect on Y that is not via D. Similarly, one cannot implement a standard covariate adjustment analysis such as regression on the trend scale, as the exposure trend $\Delta D = D_1 - D_0$ and the outcome trend $\Delta Y = Y_1 - Y_0$ is confounded by unmeasured variables $U = (U_0, U_1)$ (Figure 1b). Progress can however be made if conditional on X, the variable Z does not have any direct effect on the outcome trend except via the exposure trend, and moreover Z is independent of the unmeasured exposure-outcome confounders on the trend scale [38]. Such exogenous variable Z is referred to as an instrument for difference-in-differences. As an example, assume that one wants to assess the side effect of weight gain due to metformin versus sulfonylurea among new diabetic patients with hypertension. Here, the clinician's preference is used as an instrument. When clinicians who prefer metformine as first-line anti-diabetic therapy also more often prescribe propranolol and metoprolol (two medications that might cause weight gain) to treat hypertension, the exclusion restriction assumption in standard instrumental variable method will be violated. However, the clinician's preference may still qualify as an instrument for diference-in-differences, if the use of metformin/sulfonylurea evolves differently between the high- and low-preference groups of clinicians over time, but the use of anti-hypertensive medications that induce weight gain in the two groups does not change over time.

Quite intuitively, a valid instrument for difference-in-differences as in the above example will allow one to estimate $\beta(x)$ by a Wald-type estimator derived from the identity:

$$\beta(x) = \frac{E(Y_1 - Y_0 \mid X = x, Z = 1) - E(Y_1 - Y_0 \mid X = x, Z = 0)}{E(D_1 - D_0 \mid X = x, Z = 1) - E(D_1 - D_0 \mid X = x, Z = 0)} \tag{2.1}$$

provided that $E(D_1 - D_0 \mid X = x, Z = 1) \neq E(D_1 - D_0 \mid X = x, Z = 0)$. In the discussion below, we will show that this identification result can be obtained by viewing $\beta(x)$ as a parameter indexing an additive structural mean model for instrumented difference-in-differences. The advantage of such model is that it easily enables extensions, e.g. when the treatment effect on the multiplicative scale is of more interest. An additive structual mean model can be formally expressed as:

$$E(Y_1^{d^*} - Y_0^{d} \mid X = x, Z) = \beta(x) \times (d^* - d) + m(x), \tag{2.2}$$

for all d, d^*, where $\beta(x)$ and $m(x)$ are unknown. This model embodies the assumptions that (i) the average outcome trend given X under the same exposure over time is unchanged across strata defined by Z, and that (ii) the (time-independent) average exposure effect is constant across stratum defined by Z. In other words, Z is assumed to not modify the effect of time and of exposure on the outcome on the additive scale. Of note, Ye et al. [38] previously proposed the

assumption that $Y_1^d - Y_0^d \perp\!\!\!\perp Z \mid X$ and that $Y_t^{d^*} - Y_t^d \perp\!\!\!\perp Z \mid X$. This so-called *independence & exclusion restriction* will imply both conditions (i) and (ii).

To link $\beta(x)$ to the observed data, Ye et al. further assume that there are no unmeasured confounders of the relationship between Z and (D_t, Y_t) that simultaneously modify the effect of Z on D_t and of D_t on Y_t, given X [38]. A similar assumption is also adopted in standard instrumental variable settings to identify the average exposure effect [7, 36]. In practice, model (2.2) is more likely to hold when the unmeasured confounders U_0 and U_1 do not modify the effect of D_t on Y_t on the additive scale. While this so-called no unmeasured effect modifiers (NUEM) assumption is stronger than the assumption of no unmeasured common effect modifiers (NUCEM) proposed by Ye et al. (in the sense that NUEM holds implies that NUCEM holds but not vice versa), the former does not require knowledge on the exposure assignment at each time point.

As an example, model (2.2) holds when the outcome generating mechanism at each time point obeys the following linear models:

$$Y_1 = \alpha_1 + \beta_1 D_1 + \beta_2 D_1 X + \gamma U_0 + \delta Z + \epsilon_1,$$
$$Y_0 = \alpha_0 + \beta_1 D_0 + \beta_2 D_0 X + \gamma U_1 + \delta Z + \epsilon_0,$$

where ϵ_1 and ϵ_0 are mean-zero, normally distributed random errors (conditional on the variables in these respective models), and U_0 and U_1 are equal in distribution given X and Z. In the Supplementary Material [34], we show that under Assumption 1 and 2, $\beta(X)$ can be linked to the observed data by identification result (2.1).

Interestingly, when there are unmeasured exposure effect modifiers that lead to the violation of model (2.2), one could still identify the conditional average exposure effect among the exposed, given that some other assumptions are satisfied. When $X = x$, this effect can be expressed as: $\beta^*(x) = E(Y_t^1 - Y_t^0 \mid D_t = 1, X = x)$ for $t = 0, 1$, which is also assumed to be unchanged over time. To estimate $\beta^*(x)$, we consider the following model:

$$E(Y_t^d - Y_t^0 \mid D_t = d, X = x, Z) = \beta^*(x) \times d \quad \text{for} \quad d = 0, 1 \tag{2.3}$$

which embodies the assumption that Z itself does not modify the exposure effect among the exposed on the additive scale. Such an assumption is satisfied when $Y_t^1 - Y_t^0 \perp\!\!\!\perp Z \mid X, D_t = 1$ for $t = 0, 1$. When Z, in addition, does not modify the effect of time on the outcome in the absence of the exposure, which is satisfied when:

Assumption 3. $Y_1^0 - Y_0^0 \perp\!\!\!\perp Z \mid X,$

then $\beta^*(x)$ can be expressed as the right-hand side of expression (2.1). Notice that Assumption 3 and the conditional independence assumption $Y_t^1 - Y_t^0 \perp\!\!\!\perp Z \mid X, D_t = 1$ that imply model (2.3) are weaker than the independence and exclusion restriction (i.e. $Y_1^d - Y_0^d \perp\!\!\!\perp Z \mid X$ and $Y_t^{d^*} - Y_t^d \perp\!\!\!\perp Z \mid X$) that imply model (2.2). This results from the fact that model (2.3) aims to estimate the

exposure effect in the subgroup of exposed individuals only, while model (2.2) targets the exposure effect in the whole population, which is more challenging.

In addition to the average exposure effect among the exposed or in the whole population, other causal parameters can also be linked to the observed data by the Wald ratio on the trend scale (2.1), under some alternative assumptions. For instance, assuming monotonicity instead of NUEM or NUCEM implies that the Wald ratio (2.1) will identify a complier average exposure effect [38]. In the fuzzy difference-in-differences design, De Chaisemartin and d'Haultfoeuille [8] show that ratio (2.1) identifies the average exposure effect among patients switching from unexposed to exposed between period $t = 0$ and $t = 1$ due to the introduction of some public policy. This is under the assumption that the exposure status can only transition in one direction within each stratum of Z [8]. Meanwhile, instrumented difference-in-differences allows the exposure status of each individual to vary in any direction over time, arguably making it more appropriate to assess medical or epidemiological exposures. For instance, a patient previously exposed to a medication might become unexposed, or vice versa, due to the improvement, or worsening, of her clinical status at each time. Also, fuzzy difference-in-differences requires the percentage of exposed units in the stratum $Z = 0$ to remain constant over time (due to this group likely not affected by the policy), or else the exposure effect at time $t = 1$ among the switchers in each stratum of Z would be homogeneous. Instead, instrumented difference-in-differences requires distinct assumptions on the potential of effect modification by Z or by baseline covariates to identify the target estimand $\beta(X)$. In short, although instrumented difference-in-differences and fuzzy difference-in-differences achieve similar identification results on the additive scale, they target different types of exposure effects and consider different causal assumptions that are specific to the (public policy or healthcare) contexts for which each method is intended.

Finally, note that estimation strategies based on the identification formula (2.1) can be developed by using semi-parametric theory [38]. These strategies rely on a projection of the true function $\beta_0(X)$ of $\beta(X)$ on a parametric working model $\beta(X, \psi)$ for some finite dimensional vector ψ. Particularly, when one considers $\beta(X, \psi) = \psi$, the obtained projection will be the average treatment effect, i.e. $E\{\beta(X)\}$. Such a strategy have also been considered in standard IV settings, e.g., see the discussion by Ogburn et al. [19], or Kennedy et al. [16]. Interested readers are thus encouraged to consult these previous works for the estimation of linear structural mean models with instrumented difference-in-differences.

3. Multiplicative structural mean models for instrumented difference-in-differences

3.1. Identification

In this section, we extend the above discussion to a multiplicative structural mean model for instrumented difference-in-differences. Such an extension is po-

tentially useful for many reasons. First, discrete outcomes that are count or rare binary variables often lend themselves to log-linear models rather than linear models. In addition to that, in some clinical applications, the assumption of the instrument not modifying the effect of time and of the exposure on the outcome may be more likely to hold on the multiplicative scale rather than on the additive scale [32].

Assume that one wants to identify and estimate the average exposure effect $\beta(x)$ defined on the multiplicative scale, which is supposed to be unchanged over time, i.e. $\beta(x) = E(Y_t^1 \mid X = x)/E(Y_t^0 \mid X = x)$ for $t = 0, 1$. To achieve this, we consider the following log-linear model:

$$E(Y_1^{d^*} \mid X = x, Z) = E(Y_0^d \mid X = x, Z)e^{\beta(x)\times(d^*-d)+m(x)}. \tag{3.1}$$

Model (3.1) can be viewed as an extension of model (2.2) to the multiplicative scale. This model embodies the aforementioned assumption that Z does not modify the effect of time on the outcome on the multiplicative scale, i.e.

$$\frac{E(Y_1^d \mid X, Z = z)}{E(Y_0^d \mid X, Z = z)} = e^{m(X)} \tag{3.2}$$

nor the effect of the exposure on the outcome on the multiplicative scale, i.e.

$$\frac{E(Y_t^{d^*} \mid X, Z = z)}{E(Y_t^d \mid X, Z = z)} = e^{\beta(X)\times(d^*-d)} \tag{3.3}$$

for $z = 0, 1$ and $d, d^* = 0, 1$. As in the linear model setting, the log-linear model (3.1) is more likely to hold when the unmeasured confounders U_0 and U_1 do not modify the exposure effect at each time point, on the multiplicative scale. For instance, model (3.1) holds when the outcome generating mechanism at each time point obeys the following log-linear models:

$$E(Y_1|D_0, D_1, U_0, U_1, X, Z) = e^{\alpha_1+\beta_1 D_1+\beta_2 D_1 X+\gamma U_1+\delta Z}$$
$$E(Y_0|D_0, U_0, X, Z) = e^{\alpha_0+\beta_1 D_0+\beta_2 D_0 X+\gamma U_0+\delta Z}$$

where U_0 and U_1 are equal in distribution given X and Z. In the Supplementary Material, we prove that under Assumptions 1 and 2, $\beta(X)$ is linked to the observed data by the following moment condition:

$$E\{Y_1 e^{-\beta(X)D_1} - Y_0 e^{-\beta(X)D_0+m(X)} \mid X, Z\} = 0 \tag{3.4}$$

As for the additive structural mean model, in the presence of unmeasured exposure effect modifiers that lead to the violation of model (3.1), one can alternatively target the conditional average exposure effect among the exposed on the multiplicative scale, i.e.:

$$\beta^*(x) = \log \frac{E(Y_t^1 \mid D_t = 1, X = x)}{E(Y_t^0 \mid D_t = 1, X = x)}$$

for $t = 0, 1$. This effect is also assumed to be unchanged overtime. To estimate $\beta^*(x)$, we consider the following structural mean model:

$$E(Y_t^d \mid D_t = d, X, Z) = E(Y_t^0 \mid D_t = d, X, Z)e^{\beta^*(X)d} \tag{3.5}$$

for $d = 0, 1$ and $t = 0, 1$. This model assumes that Z does not modify the effect of the exposure on the multiplicative scale among the exposed, that is:

$$\frac{E(Y_t^1 \mid D_t = 1, X, Z = z)}{E(Y_t^0 \mid D_t = 1, X, Z = z)} = e^{\beta^*(X)}$$

Notice that this assumption is weaker than the multiplicative scale assumption 3.3 implied by model (3.1), because the former only requires no effect modification by Z in the subgroup of exposed individuals, given X.

To estimate $\beta^*(X)$, we further adopt the assumption that in the absence of exposure, Z does not modify the effect of time on the outcome on the multiplicative scale, that is:

Assumption 4. *There exist a function $m(X)$ such that:*

$$E(Y_1^0 \mid X, Z = z) = E(Y_0^0 \mid X, Z = z)e^{m(X)}.$$

This assumption is also weaker than Assumption 3.2 enclosed in model (3.1), since the latter requires no time effect modification by Z in the whole population, while the former only requires it among the exposed. Under Assumption 4, $\beta^*(X)$ can be linked to the observed data by an expression similar to (3.4).

3.2. Estimation without baseline covariates

In the specific setting where $m(X) = 0$ and $\beta(x) = \beta$, and no adjustment for observed baseline covariates is needed for the identification assumptions to hold, the target parameter e^{β} will express the average exposure effect on the multiplicative scale. The moment condition (3.4) then implies that:

$$E(Y_1 e^{-\beta D_1} \mid Z{=}1)\ E(Y_0 e^{-\beta D_0} \mid Z{=}0){=}E(Y_0 e^{-\beta D_0} \mid Z{=}1)\ E(Y_1 e^{-\beta D_1} \mid Z{=}0). \tag{3.6}$$

Solving the sample analog of this equation returns a consistent estimator $\hat{\beta}$ for β. Obtaining a closed-form expression for $\hat{\beta}$ is not possible in general cases. However, when the exposure is binary ($D_0, D_1 = 0, 1$), equation (3.6) can be rewritten in a quadratic form as:

$$\begin{aligned}(E_{111}E_{000} - E_{110}E_{001})\theta^2 - (E_{11}E_{000} + E_{111}E_{00} - E_{10}E_{001} - E_{110}E_{01})\theta \\ + E_{11}E_{00} - E_{10}E_{01} = 0,\end{aligned}$$

where $\theta = e^{-\beta} - 1$, $E_{tz} = E(Y_t \mid Z = z)$ and $E_{ttz} = E(Y_t D_t \mid Z = z)$ for $t, z = 0, 1$. The asymptotic distribution of $\hat{\beta}$ can be established using standard M-estimation theory or non-parametric bootstrap sampling.

3.3. Estimation with baseline covariates

We now discuss estimation strategies when the set of baseline covariates X is non-empty. For this, we will assume that $\beta(x) = \beta_0 + \beta_1^{\mathrm{T}} x$, but the proposed methods will work for any other finite-dimensional parametrization of $\beta(x)$, provided that this parametrization is correct. With a slight abuse of notation, we denote $\beta^{\mathrm{T}} = \begin{pmatrix}\beta_0 & \beta_1^{\mathrm{T}}\end{pmatrix}$ as the k-dimensional vector of parameters indexing $\beta(x)$.

We consider the following two settings. In the first setting, we let the covariate function $m(X)$ in the structural mean model (3.1) be correctly parametrized, in the sense that $m(X) = m(X, \gamma)$ for some finite-dimensional parameter γ. In the second setting, $m(X)$ is unspecified. In both cases, we will denote:

$$\epsilon = \epsilon(O, \beta, m(\cdot)) = Y_1 e^{-\beta(X)D_1} - Y_0 e^{-\beta(X)D_0 + m(X)}.$$

The moment condition (3.4) implies that $E(\epsilon \mid X, Z) = 0$.

Setting 1: $m(X)$ specified. Assume that $m(X, \gamma)$ is correctly specified. To construct consistent estimators for β and γ, we first note that these parameters actually index a semi-parametric model $\mathcal{M}$, represented by the class of distributions $\mathcal{P}$ of the observed data satisfying (3.4), i.e. for which $\int \epsilon(o, \beta, \gamma) d\mathcal{P}(d_0, y_0, d_1, y_1 \mid X, Z, \beta, \gamma) = 0$. From this restriction, one can derive the space of all influence functions (i.e. the orthogonal nuisance tangent space) of $\mathcal{M}$. Because of the deep connection between (asymptotically linear) estimators for a given model and the influence functions under that model, if we can find all the influence functions for $\mathcal{M}$, we can characterize all regular asymptotic linear estimators for $\mu = \begin{pmatrix}\beta^{\mathrm{T}} & \gamma^{\mathrm{T}}\end{pmatrix}^{\mathrm{T}}$ up to asymptotic equivalence [27, 15].

Theorem 1. *Suppose that Assumption 1, 2 and model (3.1) hold. The space of all influence functions for $\mu = \begin{pmatrix}\beta^{\mathrm{T}} & \gamma^{\mathrm{T}}\end{pmatrix}^{\mathrm{T}}$ under the proposed specification of $m(X, \gamma)$ in model (3.1) is $\Lambda_1^{\perp} = \left\{d^{q\times 1}(X, Z)\cdot\epsilon\right\}$, where q denotes the dimension of μ and $d^{q\times 1}(X, Z)$ is an arbitrary q-dimensional vector function of X and Z that satisfies*

$$E\left\{d^{q\times 1}(X, Z)\left(\frac{\partial \epsilon}{\partial \mu}\right)^{\mathrm{T}}\right\} = I^{q\times q}.$$

Here and below, $I^{q\times q}$ denotes the $q \times q$ identity matrix.

Theorem 1 suggests that μ can be estimated by solving the sample equivalent of the moment condition $E\{d(X, Z)\epsilon\} = 0$, where $d(X, Z)$ is an arbitrary non-trivial q-dimensional vector function of X and Z, e.g. $d^{\mathrm{T}}(X, Z) = \begin{pmatrix}1 & X^{\mathrm{T}} & Z\end{pmatrix}$. A straightforward application of the M-estimation method then allows one to derive the asymptotic variance of $\hat{\mu}$ obtained from this approach. More precisely, $\sqrt{n}(\mu - \mu_0)$ converges in distribution to:

$$N\left[0,\ E\left\{-\frac{\partial f}{\partial \mu}(O, \mu_0)\right\}\ \mathrm{var}\left\{f(O, \mu_0)\right\}\ E\left\{-\frac{\partial f}{\partial \mu}(O, \mu_0)\right\}^{-1,\mathrm{T}}\right]$$

where μ_0 denotes the true values of μ and $f(O, \mu_0) = d(X, Z)\epsilon$ denotes the estimating function. Alternatively, one can obtain the asymptotic variance of $\hat{\mu}$ by using non-parametric bootstrap sampling (Table 1).

TABLE 1
Estimation procedure when $m(X)$ is specified, i.e. $m(X) = m(X, \gamma)$.

Step	Action
1	Obtain an estimate $\hat{\mu}$ for $\mu = (\beta, \gamma)$ by solving: $$\sum_{i=1}^{n} d^{q\times 1}(X, Z)[Y_{1i}e^{-\beta(X_i)D_{1i}} - Y_{0i}e^{-\beta(X_i)D_{0i}+m(X_i)}] = 0^{q\times 1},$$ where $d^{q\times 1}(X, Z)$ is an arbitrary $q \times 1$ function of (X, Z), and q is the dimension of μ.
2	Estimate the variance of $\hat{\mu}$ by M-estimation or bootstrap.

For completeness, we also derive in the Supplementary Material the efficient influence function among the elements of $\Lambda_1^{\perp}$, by projecting the score of θ (under the true parametric submodel) on $\Lambda_1^{\perp}$. However, we do not recommend the use of this efficient influence function in practice. First, it involves $\text{var}(\epsilon \mid X, Z)^{-1} = E^{-1}(\epsilon^2 \mid X, Z)$ as a nuisance parameter. The efficiency of the resulting estimator is thus local in the sense that it is only attained when this variance can be estimated consistently at sufficiently fast rates. Even when a proper estimate can be obtained for $\text{var}(\epsilon \mid X, Z)$, the inverse of this variance can make the resulting estimator for θ become very unstable, which makes it difficult to perform well in practice.

Setting 2: $m(X)$ unspecified. We now discuss a more general setting where the function $m(X)$ in model (3.1) is left unspecified. For this, consider first an easier case where $m(X)$ is a priori known. By a similar proof as in Theorem 1, one can show that under Assumption 1 and 2, the orthogonal complement of the nuisance tangent space in model (3.1) (given that it is correctly specified) is $\Lambda_1^{\perp} = \{d^{k\times 1}(X, Z)\epsilon\}$, where $d^{k\times 1}(X, Z)$ is arbitrary but satisfies:

$$E\left\{d^{k\times 1}(X, Z)\left(\frac{\partial \epsilon}{\partial \beta}\right)^{\mathrm{T}}\right\} = I^{k\times k}.$$

To recognize that $m(X)$ is unknown, we then need to determine the subspace of mean-zero functions in $\Lambda_1^{\perp}$ that is additionally orthogonal to the nuisance scores for $m(X)$.

Theorem 2. *Suppose that Assumptions 1, 2 and model (3.1) hold. The orthogonal complement of the nuisance tangent space of model (3.1) when $m(X)$ is left unspecified is $\Lambda_2^{\perp} = \{[d^{k\times 1}(X, Z) - d^{*,k\times 1}(X, Z)]\epsilon\}$, where $d^{k\times 1}(X, Z)$ is an arbitrary k-dimensional function of X and Z that satisfies*

$$E\left\{[d^{k\times 1}(X, Z) - d^{*,k\times 1}(X, Z)]\left(\frac{\partial \epsilon}{\partial \beta}\right)^{\mathrm{T}}\right\} = I^{k\times k}$$

with:

$$d^{*,k\times 1}(X,Z)=\frac{\lambda(X,Z)\sigma^{-2}(X,Z)}{E\{\lambda^2(X,Z)\sigma^{-2}(X,Z)\mid X\}}E\{d^{k\times 1}(X,Z)\lambda(X,Z)\mid X\}$$

and $\lambda(X,Z)=E(Y_0e^{-\beta(X)D_0}|Z,X)/E(Y_0e^{-\beta(X)D_0}|X)$.

A direct consequence of Theorem 2 is that elements in $\Lambda_2^{\perp}$ have mean zero even when $m(X)$ is misspecified, i.e. $m(X)\neq m_0(X)$, where $m_0(X)$ denotes the true form of $m(X)$ that is unknown, provided that $\sigma^{-2}(X,Z)$, $\lambda(X,Z)$, $E\{d(X,Z)\lambda(X,Z)\mid X\}$ and $E(\lambda^2(X,Z)\sigma^{-2}(X,Z)\mid X)$ are correctly specified. Note that postulating parametric models for these nuisance parameters is not entirely satisfactory, as it may easily lead to model misspecification and incompatibility. Furthermore, the estimating functions in $\Lambda_2^{\perp}$ are highly complex (e.g. due to the presence of many complicated nuisance parameters), which may lead to convergence issues in practice.

To remedy this, consider the element ν of $\Lambda_2^{\perp}$ corresponding to $d^{k\times 1}(X,Z)=g^{k\times 1}(X,Z)-E\{g^{k\times 1}(X,Z)\lambda(X,Z)\mid X\}$, where $g^{k\times 1}(X,Z)$ is an arbitrary k-dimensional vector function of X and Z satisfying the conditions in Theorem 2. This element ν can be alternatively expressed as:

$$\nu=\{g(X,Z)-E(g(X,Z)\lambda(X,Z)\mid X)\}\{Y_1e^{-\beta(X)D_1}-Y_0e^{-\beta(X)D_0+m(X)}\},$$

Theorem 2 then implies that $E(\nu)=0$ when $m(X)\neq m_0(X)$, given that only $\lambda(X,Z)$ and $E(g(X,Z)\lambda(X,Z)\mid X)$ are consistently estimated. In our current setting with binary exposure, by fixing $m(X)=0$, the moment condition $E(\nu)=0$ can be reexpressed as:

$$E\Bigg[\underbrace{\left\{g(X,Z)-\frac{(e^{-\beta(X)}-1)a_1(X)+a_2(X)}{(e^{-\beta(X)}-1)a_3(X)+a_4(X)}\right\}}_{A(O)}\times\underbrace{\{(Y_1D_1-Y_0D_0)(e^{-\beta(X)}-1)+Y_1-Y_0\}}_{B(O)}\Bigg]=0 \tag{3.7}$$

Here, we denote $a_1(X)=E\{Y_0D_0g(X,Z)\mid X\}$, $a_2(X)=E\{Y_0g(X,Z)\mid X\}$, $a_3(X)=E\{Y_0D_0\mid X\}$ and $a_4(X)=E\{Y_0\mid X\}$.

We now construct an estimation strategy for the parameter vector $\beta^{k\times 1}$ based on the moment condition (3.7). Since $\beta^{k\times 1}$ is the zero of this moment condition, it can be viewed as a well defined model-free population parameter without reference to the original model (3.1). This suggests that one can work under the non-parametric model and estimate $\beta^{k\times 1}$ by using semi-parametric theory, to enable fast $\sqrt{n}$ rates of convergence to the parameter of interest. This is achievable even when nuisance functions $a_1(X)$ to $a_4(X)$ are estimated at slower rates, e.g. using flexible data-adaptive or machine learning methods.

In what follows, we will focus on the special case where $\beta(X)=\beta$ (i.e. X are not effect modifiers and $k=1$) for the sake of simplicity. In the Supplementary

Material, we generalize the discussion to more general cases where $\beta(X)$ is a non-constant parametric function of X (i.e. $k > 1$). Denote $\theta = e^{-\beta} - 1$. To characterize its influence function under the non-parametric model (so as to obtain a non-parametric estimator), we first rewrite θ as $\theta = \theta(\mathcal{P})$ to stress that θ is a functional of the observed data distribution $\mathcal{P}$. In what follows, we perturb θ in the direction $\tilde{\mathcal{P}}_t$ of a point mass at single observation $\tilde{o}$ of O, i.e. $\tilde{\mathcal{P}}_t = (1-t)\mathcal{P} + t\mathbb{1}_{\tilde{o}}$ where $\mathbb{1}_{\tilde{o}}$ denotes the Dirac delta function at $\tilde{\boldsymbol{o}}$. The efficient influence function of θ at observation $\tilde{o}$ under the non-parametric model for the observed data can then be identified by evaluating the Gateaux derivative of $\theta(\mathcal{P}_t)$ with respect to t at $t = 0$, that is

$$\phi(\tilde{o}, \theta, \eta) = \left.\frac{d\theta(\mathcal{P}_t)}{dt}\right|_{t=0}$$

where $\eta = c(a_1, a_2, a_3, a_4, a_5, a_6)$ is the vector of all nuisance parameters. For a more detailed guidance on influence functions, see Hines et al. [12] and Kennedy [15].

Proposition 1. *Under certain regularity conditions, it can be shown that the influence function of θ under the non-parametric model is:*

$$\begin{aligned}\phi(\boldsymbol{O}, \theta, \eta) = {} & -C^{-1}A(O)B(O) + C^{-1}\times \\ & \times \left[\left\{\frac{\theta Y_0 g(X,Z)(D_0+1) - a_1\theta - a_2}{\theta a_3 + a_4}\right.\right. \\ & \left.\left. - \frac{(\theta a_1 + a_2)(\theta Y_0 D_0 - \theta a_3 + Y_0 - a_4)}{(\theta a_3 + a_4)^2}\right\} E\big\{B(O) \mid X\big\}\right]\end{aligned}$$

where $C = C(O, \theta, \eta) = E\big\{(a_2 a_3 - a_1 a_4)(\theta a_3 + a_4)^{-2} B(O) + A(O)(Y_1 D_1 - Y_0 D_0)\big\}$.

As $E\{\phi(O, \theta, \eta)\} = 0$ by construction, one can obtain an estimate $\hat{\theta}$ by solving the sample analog of this equation, $\sum_i^n \phi(O_i, \hat{\theta}, \hat{\eta}) = 0$, where $\hat{\eta} = (\hat{a}_1, \hat{a}_2, \hat{a}_3, \hat{a}_4, \hat{a}_5, \hat{a}_6)$ denotes an estimate for η, possibly obtained by flexible data-adaptive or machine learning methods. Assume that $\hat{\eta}$ converges in probability to some η_1 that might be potentially different from the true value of the nuisance parameter η. In the Supplementary Material, we prove that the remainder term: $R(\eta, \eta_1) := \theta(\eta_1) - \theta(\eta) + E\{\phi(O, \eta_1)\}$ is a second order term involving only products of the type $E[c(\eta, \eta_1)\{f(\eta_1) - f(\eta)\}\{g(\eta_1) - g(\eta)\}]$. This result will be useful when establishing the asymptotic properties of $\hat{\theta}$, as shown in the theorem below.

Theorem 3 (Asymptotic normality and efficiency)**.** *Assume that (i) the second-order term $R(\hat{\eta}, \eta)$ is $o_P(n^{-1/2})$ and (ii) the class of functions $\{\phi(\eta, \theta') : |\theta' - \theta| < \delta, ||\eta - \eta_1|| < \delta\}$ is Donsker for some $\delta > 0$ and such that $\mathrm{pr}\{\phi(\eta, \theta') - \phi(\eta_1, \theta)\}^2 \to 0$ as $(\eta, \theta') \to (\eta_1, \theta)$, then: $\hat{\theta}(\hat{\eta}) = \theta(\eta) + \frac{1}{n}\sum_{i=1}^n \phi(O_i, \eta) + o_P(1/\sqrt{n})$, due to which $\sqrt{n}(\hat{\theta}_1 - \theta)$ converges in distribution to $N(0, \zeta^2)$, where $\zeta^2 = \mathrm{var}\{\phi(O, \eta)\}$ is the non-parametric efficiency bound.*

The proof of this theorem follows the general proof presented in Chernozhukov et al. [5]. Some remarks are noteworthy here. First, condition (i) for asymptotic normality in Theorem 3 requires that all components of $\hat{\eta}$ converges in $L_2(P)$ norm to their true counterparts in η at faster than $n^{1/4}$-rate, to ensure that the remainder term $R(\hat{\eta}, \eta)$ is of second order (see above). Under certain conditions, this can be satisfied by data-adaptive algorithms such as regression trees [35], neural networks [4], and highly adaptive lasso [29].

Condition (ii) (i.e. Donsker condition) restricts the flexibility of the nuisance estimators, but Donsker classes still cover many complex functions such as Lipschitz functions and so forth [31, 15]. Alternatively, one can avoid condition (ii) by using cross-fitting in the estimation procedure, given that the nuisance estimators $\hat{\eta}$ are consistent and satisfy condition (i). The exact steps of this procedure are described in Table 2.

A direct consequence of Theorem 3 is that the (asymptotic) behavior of the proposed estimator $\hat{\theta}$ is the same as if the nuisance parameters η were known. As such, one can quite easily obtain a sandwich estimator of the asymptotic variance of $\hat{\theta}$ as the sample variance of $\phi(O, \hat{\eta})$. This variance estimate may be used to construct Wald-type confidence intervals.

TABLE 2
Estimation procedure when $m(X)$ is unspecified and $\beta(X) = \beta$.

Step	Action
1	Partition the index set $\{1, \ldots, n\}$ into Q set of approximately similar size, i.e. $V_1, \ldots, V_Q$.
2	For each $q = 1, \ldots, Q$: • Obtain an estimator $\hat{\eta}^q$ for η by training the prediction algorithm using data in the the sample $T_q = \{1, \ldots, n\} \setminus V_q$. • Predict $\eta(O_i)$ by $\hat{\eta}^q(O_i)$ for each unit $i \in V_q$
3	Obtain the estimate $\hat{\theta}$ for $\theta = e^{-\beta} - 1$ by solving the equation: $\sum_{i=1}^{n} \phi\{O_i, \hat{\theta}, \hat{\eta}^q\} = 0$
4	Estimate the variance of $\hat{\theta}$ by the sample variance of $\phi(O_i, \hat{\theta}, \hat{\eta}^q)$.

4. A simulation study

In this section, we conduct a simulation study to assess the finite sample performance of the proposed approaches. The aim of the analysis is to estimate the average causal effect of a binary exposure on a count outcome (setting 1 and 3), or on a rare binary outcome with frequency around 10–12% (setting 2 and 4). Assume that each patient is followed up over two time points (i.e. longitudinal data structure). At each time point, the exposure-outcome relationship is confounded by an unmeasured variable U_t $(t = 0, 1)$. In settings 3 and 4, adjusting for a baseline covariate X is needed for the binary instrument for difference-in-differences Z to be valid. The average treatment effect is $\beta(X) = 0$ across all settings. Other details about the data generating mechanism are provided in Table 3.

TABLE 3
Simulation study: Data generating mechanism.

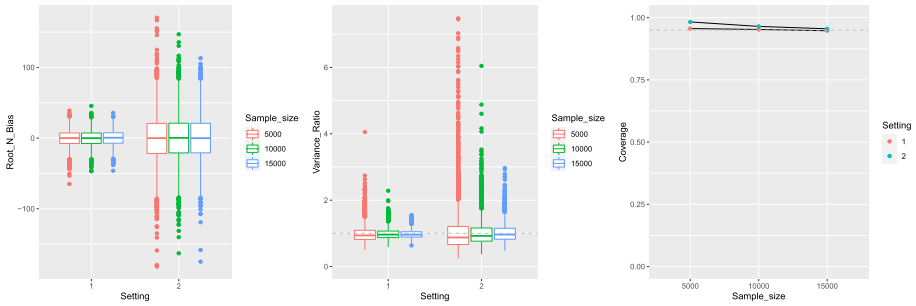
Setting	Characteristics	Data generating mechanism
1	Baseline	$Z \sim B(N, 0.5)$
	Time $t = 0$	$U_0 \sim N(0.5, 1)$
		$E(D_0 \mid U_0, Z) = \text{expit}(1 - Z + U_0)$
		$E(Y_0 \mid Z, U_0, D_0) = \exp(-1 + 0D_0 + 0.5U_0 + 0.5Z)$
	Time $t = 1$	$U_1 \sim N(0.5, 1)$
		$E(D_1 \mid U_1, U_1, Z) = \text{expit}(-1 + Y_0 + U_1 + Z)$
		$E(Y_1 \mid Z, U_1, D_1, Y_0) = \exp(-1 + 0D_1 + 0.5U_1 + 0.5Z)$
2	Baseline	$Z \sim B(N, 0.5)$
	Time $t = 0$	$U_0 \sim U(0, 1)$
		$E(D_0 \mid X, U_0, Z) = \text{expit}(-0.85 - Z + U_0)$
		$E(Y_0 \mid X, Z, U_0, D_0) = \exp(-3.7 + 0D_0 + U_0 + Z)$
	Time $t = 1$	$U_1 \sim U(0, 1)$
		$E(D_1 \mid X, U_0, U_1, Z) = \text{expit}(0.272 + Y_0 + U_1 + Z)$
		$E(Y_1 \mid X, Z, U_1, D_1, Y_0) = \exp(-3.9 + 0D_1 + U_1 + Z)$
3	Baseline	$X = \min\{P(0.5) + 0.5, 2.5\}$
		$E(Z \mid X) = \text{expit}(-0.5 + X)$
	Time $t = 0$	$U_0 \sim N(0.5, 1)$
		$E(D_0 \mid U_0, Z) = \text{expit}(1 - Z + U_0 + X_0)$
		$E(Y_0 \mid Z, U_0, D_0) = \exp[-1 + 0D_0 + 0.5U_0 + 0.5Z + 0.25X + 0.15\sin(X)]$
	Time $t = 1$	$U_1 \sim N(0.5, 1)$
		$E(D_1 \mid U_1, U_1, Z) = \text{expit}(-1 + Z + U_1 + Y_0 + X)$
		$E(Y_1 \mid Z, U_1, D_1, Y_0) = \exp[-1 + 0D_1 + 0.5U_1 + 0.5Z + 0.35X + 1.70\sin(X)]$
4	Baseline	$X = \min\{P(0.5) + 0.5, 3.5\}$
		$E(Z \mid X) = \text{expit}(-0.8 + X)$
	Time $t = 0$	$U_0 \sim U(0, 1)$
		$E(D_0 \mid X, U_0, Z) = \text{expit}(-0.85 - Z + U_0 + X)$
		$E(Y_0 \mid X, Z, U_0, D_0) = \exp[-1.8 + 0D_0 - 1.5U_0 - 0.25Z + 0.15X + 0.15\sin(X)]$
	Time $t = 1$	$U_1 \sim U(0, 1)$
		$E(D_1 \mid X, U_0, U_1, Z) = \text{expit}(0.272 + Y_0 + 0.5U_1 + 0.5Z + 0.5X)$
		$E(Y_1 \mid X, Z, U_1, D_1, Y_0) = \exp[-3 + 0D_1 - 1.5U_1 - 0.25Z + 0.35X + 1.70\sin(X)]$

Across all settings, using Z as a standard instrument will return a biased estimate for β as the exclusion restriction assumption is violated (i.e. Z has a direct effect on Y_t that does not go through D_t). The instrumented difference-in-differences approach is alternatively used to analyze the data as follow:

In settings 1 and 2 with no observed covariates, we estimate β by solving equation (3.6).

In settings 3 and 4, the function $m(X)$ in the underlying structural mean model (3.1) has the form $m(X) = \beta_0 + \beta_1 X + \beta_2 \sin(X)$. We consider three approaches to estimate β.

- **Approach (A1)**: $m(X)$ is mis-specified as $m(X) = \delta_0 + \delta_1 X$. The parameter vector $\theta = \begin{pmatrix}\delta_0 & \beta & \delta\end{pmatrix}^{\mathrm{T}}$ is estimated by solving the equation $\sum_{i=1}^{N} d(X, Z)\epsilon = 0$, where $d(X, Z) = \begin{pmatrix}1 & X & Z\end{pmatrix}^{\mathrm{T}}$.
- **Approach (A2)**: $m(X)$ is correctly specified as shown above. The parameter vector $\theta = \begin{pmatrix}\beta_0 & \beta & \beta_1 & \beta_2\end{pmatrix}^{\mathrm{T}}$ is estimated by solving the equation $\sum_{i=1}^{N} d(X, Z)\epsilon = 0$, where $d(X, Z) = \begin{pmatrix}1 & X & Z & \sin(X)\end{pmatrix}^{\mathrm{T}}$.
- **Approach (A3)**: $m(X)$ is unspecified. β is estimated by the non-parametric approach discussed in Section 3.3. The nuisance parameters involved in this approach are estimated by using the super learner algorithm [30], whose library includes the main terms generalized linear model, the multivariate adaptive regression splines and the highly adaptive lasso.



(a) Root-N Bias (b) Variance Ratio (c) Coverage

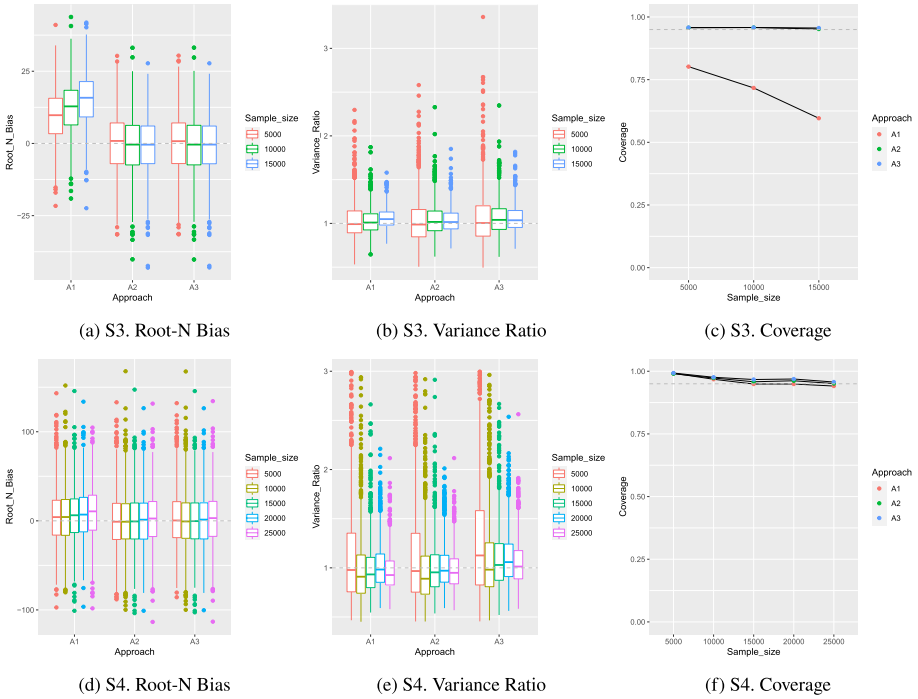
FIG 2. *Simulation results – Setting 1 and 2. (a): the distribution of* $\sqrt{n}$*-bias, i.e.* $\sqrt{n}(\hat{\beta}_i - \beta)$*, where* $\hat{\beta}_i$ *denotes the estimate of* β *obtained from simulation* i*; (b): the distribution of the ratio between the variance estimate* $\hat{V}(\hat{\beta}_i)$ *and the true variance* $V(\hat{\beta})$*; (c): coverage of the 95% Wald confidence interval for* β*.*

Although cross fitting is required to ensure valid inference for approach A3 without relying on the Donsker condition, we will not consider it here to shorten the computational time of the simulation study.

Three sample sizes, $n = \{5, 10, 15\} \times 10^3$, are considered in each setting. In setting 4 (rare binary outcome with observed baseline covariates), two other sample sizes of $n = \{20, 25\} \times 10^3$ are additionally considered to further evaluate the asymptotic properties of the proposed approaches. In each setting, we assess (i) the $\sqrt{n}$-consistency of the obtained estimator $\hat{\beta}$ for β, (ii) the ratio between the variance estimate of $\hat{\beta}$ and the true variance of $\hat{\beta}$ (calculated across all simulations), and (iii) the coverage of the 95% Wald confidence interval for β. We implement 10^3 simulations in each setting.

Results of this simulation study are visualized in Figures 2 and 3. Numerical data to reproduce these figures are also provided in the Supplementary Material. When X is empty and $m(X) = 0$ (settings 1 and 2), the proposed method returns a valid estimate $\hat{\beta}$ for β that is $\sqrt{n}$-consistent (Figure 2a). In setting 2 (rare binary outcome), the variance of $\hat{\beta}$ is slightly underestimated when the sample size is small (Figure 2b). This results in a (slight) over-coverage of the 95% CI (Figure 2c). Such a problem, however, disappears when the sample size is sufficiently large ($n = 15 \times 10^3$).

When X is non-empty and $m(X) \neq 0$ (settings 3 and 4), the estimation approach based on Theorem 1 only provides a valid estimate for β (i.e. $\sqrt{n}$-consistent) when the function $m(X)$ is correctly specified (i.e. approach A2). In contrast, the non-parametric approach A3 can obtain a valid estimate for β without having to specify $m(X)$ (Figure 3a). When the outcome is a rare binary variable (setting 4), the performance of both approaches can be worsened if the sample size is insufficiently large (Figure 3b–c).



(a) S3. Root-N Bias (b) S3. Variance Ratio (c) S3. Coverage

(d) S4. Root-N Bias (e) S4. Variance Ratio (f) S4. Coverage

FIG 3. *Simulation results – Setting 3 (S3) and setting 4 (S4). (a) and (d): the distribution of* $\sqrt{n}$*-bias; (b) and (e): the distribution of the ratio between the variance estimate and the true variance of* β*. (c) and (f): coverage of the 95% Wald confidence interval for* β*.*

5. Extension to repeated cross-sectional data structure

Thus far, we have discussed the instrumented difference-in-differences method for longitudinal or panel data, in which each individual is followed-up over two time points. In this section, we extend the above results to the repeated cross-sectional, or "pseudo-longitudinal" data structure [20]. In this setting, D_t and Y_t are evaluated on an independent sample at each time point t. For instance, respondents of an annual survey in one year are different from those in the prior year. It is thus commonly assumed that there is no overlap in the samples between different periods [20]. To formalize this, denote $O^* = (Z, X, D, Y, T)$ the observed data of each individual in a repeated cross-sectional study, where $T = 0, 1$ denotes the time point, $Y = Y_1 I(T = 1) + Y_0 I(T = 0)$ and $D = D_1 I(T = 1) + D_0 I(T = 0)$. For every stratum defined by levels of Z and X, the collected data at each time point is a random sample from the population of interest, that is:

Assumption 5. $T \perp\!\!\!\perp (Y_1, Y_0, D_1, D_0) \mid X, Z$.

When X is empty, $m(X) = 0$ and $\beta(X) = \beta$, Assumption 5 implies that:

$$E(Ye^{-\beta D} \mid T = 1, Z = 1) = E(Ye^{-\beta D} \mid T = 0, Z = 1) \frac{E(Ye^{-\beta D} \mid T = 1, Z = 0)}{E(Ye^{-\beta D} \mid T = 0, Z = 0)} \tag{5.1}$$

Solving the sample analog of this equation will return a consistent estimator $\hat{\beta}$ for β. A simple application of the Delta method allows one to establish the asymptotic properties of $\hat{\beta}$.

Consider now the structural mean model (3.1) with X being non-empty and $m(X) = m(X, \gamma)$ correctly parameterized by some finite-dimensional parameter γ. To identify the orthogonal nuisance tangent space $\Lambda_3^{\perp}$ of model (3.1) under the repeated cross-sectional data structure O^*, one need to map the elements in $\Lambda_1^{\perp}$ to those in $\Lambda_3^{\perp}$. For this, note that for every $U_{O^*} \in \Lambda_3^{\perp}$, the mean of U_{O^*} given O (calculated with respect to the true distribution $\mathcal{P}$ of the full data structure O) must equal some element $U_O \in \Lambda_1^{\perp}$. The same remark also allows one to establish the orthogonal nuisance tangent space $\Lambda_4^{\perp}$ of model (3.1) in the repeated cross-sectional setting, when $m(X)$ is left unspecified.

Theorem 4. *Suppose that Assumptions 1, 2, 5 and model (3.1) hold. When the observed data structure is $O^* = (Z, X, D, Y, T)$, the orthogonal complement of the nuisance tangent space of model (3.1) under the parametrization $m(X) = m(X, \gamma)$ is:*

$$\Lambda_3^{\perp} = \left\{ d^{q\times 1}(X, Z) \cdot \pi(O, \theta) + s^{q\times 1}(X, Z) \cdot \left[T - \text{pr}(T = 1 \mid Z, X) \right] \right\},$$

where

$$\pi(O, \theta) = \frac{TYe^{-\beta(X)D}}{\text{pr}(T = 1|Z, X)} - \frac{(1 - T)Ye^{-\beta(X)D+m(X)}}{1 - \text{pr}(T = 1 \mid Z, X)}$$

and $d^{q\times 1}(X, Z)$ and $s^{q\times 1}(X, Z)$ are arbitrary q-dimensional vector functions of X and Z that satisfies

$$E\left\{ d^{q\times 1}(X, Z) \left(\frac{\partial \pi(O, \theta)}{\partial \theta} \right)^{\text{T}} \right\} = I^{q\times q}.$$

Here, q denotes the dimension of the parameter vector $\theta = \begin{pmatrix} \beta^{\text{T}} & \gamma^{\text{T}} \end{pmatrix}^{\text{T}}$. In contrast, the orthogonal complement of the nuisance tangent space of model (3.1) under the data structure O^*, when $m(X)$ is left unspecified is:

$$\Lambda_4^{\perp} = \left\{ \left[d^{k\times 1}(X, Z) - d^{*,k\times 1}(X, Z) \right] \cdot \pi(O, \beta) + q^{k\times 1}(X, Z) \cdot \left[T - \text{pr}(T = 1 \mid Z, X) \right] \right\},$$

where $d^{k\times 1}(X, Z)$ and $q^{k\times 1}(Z, X)$ are arbitrary k-dimensional functions of X and Z that satisfy:

$$E\left\{ \left[d^{k\times 1}(X, Z) - d^{*,k\times 1}(X, Z) \right] \frac{\partial \pi(O, \beta)}{\partial \beta} \right\} = I^{k\times k}$$

and $d^{*,k\times 1}(X, Z)$ is defined as in Theorem 2, but with:

$$\lambda(X, Z) = \frac{E\left[(1 - T)Ye^{-\beta(X)D} \mid Z, X \right] / \text{pr}(T = 0 \mid X, Z)}{E\left[(1 - T)Ye^{-\beta(X)D} \mid X \right] / \text{pr}(T = 0 \mid X)}.$$

A consequence of Theorem 4 is that to estimate the parameters indexing model (3.1), one needs to additionally model the nuisance parameter $\text{pr}(T = 1|X, Z)$. The estimation strategies that we have previously discussed in Section 3 can then be easily extended to this setting. Details on this are thus omitted.

6. Application to antihyperglycemic drugs on weight gain

We now apply our proposed methods to investigate the risk of moderate to severe weight gain of metformin versus sulfonylureas as initial therapy for new users of antihyperglycemic drugs (i.e. prescribed for patients with diabetes) during the period of 1995 to 2011. The data for this analysis were extracted from The Health Improvement Network [18]. From this database, we select patients who were present for at least 180 days before receiving any antihyperglycemic drugs, and then were started on an initial therapy with either metformin ($D = 1$) or a sulfonylurea ($D = 0$), with a baseline glycosylated hemoglobin (HbA1c) of $\geq 7\%$ [9]. The outcome of interest Y is a binary variable, which indicates an increase of at least 10% of BMI at two years of follow-up compared to each patient's baseline. This cut-off value is chosen based on the definition of moderate-to-severe weight gain (i.e. Grade 2–3) of the common terminology criteria for adverse events [25]. Although the cut-offs are proposed for weight, we use the same thresholds for BMI as this measure is a linear function of weight. The frequency of the outcome among patients treated with metformin and sulfonylurea is 3.6% and 11.7%, respectively.

During the research period, the use of metformin rose very quickly, while the use of sulfonylureas declined quite dramatically. Beginning in 2000, metformin became more commonly used than sulfonylurea. We thus choose the time point $t = 0$ to be the period of 1995 to 1999 (i.e. sulfonylurea more commonly used), and $t = 1$ to be the period of 2000 to 2011 (i.e. metformin more commonly used). We make the assumption that at each time point, a random sample of patients was taken from the population of interest (i.e. Assumption 4). The aforementioned variability in the prescription trends of both drugs also led us to define our instrument for difference-in-differences based on provider preference. For this, we first calculated the proportion of patients starting on metformin within each general practitioner practice in 1995. We then assigned $Z = 1$ if this proportion is larger than the median of all practices and $Z = 0$ otherwise. We did not consider baseline covariate adjustment in this analysis.

TABLE 4
Data characteristics across two timepoints.

Characteristics	**Time 0**	**Time 1**
Number of patients	1656	15234
$P(D_t = 1)$	0.46	0.86
$P(Z = 1\|T = t)$	0.58	0.53
$P(Y_t = 1\|D_t = 1)$	0.03	0.04
$P(Y_t = 1\|D_t = 0)$	0.10	0.12

Data from 16890 patients (117 practices) are finally included. By solving the sample analog of equation (5.1), we obtain an estimate of of $\beta = -1.27$ for the

treatment effect on the log relative risk scale, with a 95% confidence interval ranging from −3.07 to 0.53. This suggests that the risk of moderate to serious weight gain from metformin is $e^{-1.27} = 0.281$ times as low compared with sulfonylurea. Although this finding is not statistically significant, the direction of the result agrees with prior findings, which also suggests an increase risk of weight gain by sulfonylurea compared to other oral antihyperglycemic drugs [21, 6]. Here we focus though on the incidence of moderate to severe weight gain.

7. Conclusion

Instrumented difference-in-differences is a new addition to the range of approaches for improving causal inference in the evaluation of interventions and exposures when a randomised trial is impractical [14, 1, 37, 9, 2, 3]. In this paper, we have proposed novel additive and multiplicative structural mean models for the instrumented difference-in-differences design. By applying semi-parametric theory, we also develop multiple estimation approaches for the parameters indexing such models, thereby enabling the estimation of the average exposure effect in the whole population or among the exposed, on the additive and multiplicative scales. In the special case where the outcome indicates a rare event with a small success probability (i.e. around 10% or less as a rule of thumb), the multiplicative structural mean models can also be good approximations for the true structural mean models that have a logistic link function. However, the estimation of the treatment effect in this setting often requires a quite large sample size to obtain valid inference.

A potential direction for future research is to develop estimation strategies when a working model, defined by a least squares projection, is used to summarize $\beta(X)$. This aims to further relax the parametric assumption made on $\beta(X)$, which has been considered in the additive setting [38]. In the current multiplicative setting, one important challenge is that we do not have a closed-form expression of $\beta(X)$. Instead, $\beta(X)$ is linked to the observed data by a moment equation when all causal assumptions are satisfied. The possibilities of projecting $\beta(X)$ on a parametric working model in the absence of its closed-form expression will be further explored in future research. It is also important to develop estimation strategies for structural mean models with a logistic link function, without having to assume the binary outcome is rare. The difficulty with constructing consistent estimators for such logistic models is in finding a residual $\epsilon(O, \beta)$ satisfying a moment condition similar to (3.4), i.e. $E\{\epsilon(O, \beta) \mid X, Z\} = 0$. Extension to a logistic link may thus require a rather different line of thinking. Finally, while we here focus on two time points, the proposed models should also be extended to multiple time points settings where many additional complications may also present [24].

Supplementary Material

Supplement for "Structural mean models for instrumented difference-in-differences"
(doi: 10.1214/24-EJS2313SUPP; .pdf). Supplement A: Identification results. Proofs of the identification results. Supplement B: Estimation strategies. Proofs of the estimating strategies proposed in Section 3

References

[1] Abadie, A., Diamond, A. and Hainmueller, J. (2015). Comparative politics and the synthetic control method. *American Journal of Political Science* **59** 495–510.

[2] Baiocchi, M., Cheng, J. and Small, D. S. (2014). Instrumental variable methods for causal inference. *Statistics in Medicine* **33** 2297–2340. MR3257582

[3] Callaway, B. and Karami, S. (2023). Treatment effects in interactive fixed effects models with a small number of time periods. *Journal of Econometrics* **233** 184–208. MR4554732

[4] Chen, X. and White, H. (1999). Improved rates and asymptotic normality for nonparametric neural network estimators. *IEEE Transactions on Information Theory* **45** 682–691. MR1677026

[5] Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C. and Newey, W. (2017). Double/debiased/neyman machine learning of treatment effects. *American Economic Review* **107** 261–65.

[6] Confederat, L., Stefan, R., Lupaccu, F., Constantin, S., Avram, I., Doloca, A. and Profire, L. (2016). Side effects induced by hypoglycaemic sulfonylureas to diabetic patients-a retrospective study. *Farmacia* **64** 674–679.

[7] Cui, Y. and Tchetgen Tchetgen, E. (2021). A semiparametric instrumental variable approach to optimal treatment regimes under endogeneity. *Journal of the American Statistical Association* **116** 162–173. MR4227683

[8] De Chaisemartin, C. and d'Haultfoeuille, X. (2018). Fuzzy differences-in-differences. *The Review of Economic Studies* **85** 999–1028.

[9] Ertefaie, A., Small, D. S., Flory, J. H. and Hennessy, S. (2017). A tutorial on the use of instrumental variables in pharmacoepidemiology. *Pharmacoepidemiology and Drug Safety* **26** 357–367.

[10] Hernán, M. A., Brumback, B. and Robins, J. M. (2001). Marginal structural models to estimate the joint causal effect of nonrandomized treatments. *Journal of the American Statistical Association* **96** 440–448. MR1939347

[11] Hernán, M. A. and Robins, J. M. (2006). Instruments for causal inference: an epidemiologist's dream? *Epidemiology* 360–372.

[12] Hines, O., Dukes, O., Diaz-Ordaz, K. and Vansteelandt, S. (2022). Demystifying statistical learning based on efficient influence functions. *The American Statistician* 1–13. MR4453533

[13] Imai, K., Keele, L. and Tingley, D. (2010). A general approach to causal mediation analysis. *Psychological Methods* **15** 309.
[14] Imbens, G. W. and Lemieux, T. (2008). Regression discontinuity designs: A guide to practice. *Journal of Econometrics* **142** 615–635. MR2416821
[15] Kennedy, E. H. (2016). Semiparametric theory and empirical processes in causal inference. In *Statistical Causal Inferences and Their Applications in Public Health Research* 141–167. Springer. MR3617956
[16] Kennedy, E. H., Lorch, S. and Small, D. S. (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. MR3904782
[17] Leite, M. L. C., Nicolosi, A., Osella, A. R., Molinari, S., Cozzolino, E., Velati, C., Lazzarin, A. and Study, N. I. S. D. A. (1995). Modeling incidence rate ratio and rate difference: additivity or multiplicativity of human immunodeficiency virus parenteral and sexual transmission among intravenous drug users. *American Journal of Epidemiology* **141** 16–24.
[18] Lewis, J. D., Schinnar, R., Bilker, W. B., Wang, X. and Strom, B. L. (2007). Validation studies of the health improvement network (THIN) database for pharmacoepidemiology research. *Pharmacoepidemiology and Drug Safety* **16** 393–401.
[19] Ogburn, E. L., Rotnitzky, A. and Robins, J. M. (2015). Doubly robust estimation of the local average treatment effect curve. *Journal of the Royal Statistical Society Series B: Statistical Methodology* **77** 373–396. MR3310531
[20] Pan, X. (2022). Repeated cross-sectional design. In *Encyclopedia of Gerontology and Population Aging* 4246–4250. Springer.
[21] Phung, O. J., Scholle, J. M., Talwar, M. and Coleman, C. I. (2010). Effect of noninsulin antidiabetic drugs added to metformin therapy on glycemic control, weight gain, and hypoglycemia in type 2 diabetes. *Jama* **303** 1410–1418.
[22] Robins, J. M. (1994). Correcting for non-compliance in randomized trials using structural nested mean models. *Communications in Statistics – Theory and Methods* **23** 2379–2412. MR1293185
[23] Robins, J. M. and Tsiatis, A. A. (1991). Correcting for non-compliance in randomized trials using rank preserving structural failure time models. *Communications in Statistics – Theory and Methods* **20** 2609–2631. MR1144866
[24] Roth, J., Sant'Anna, P. H., Bilinski, A. and Poe, J. (2022). What's trending in difference-in-differences? A synthesis of the recent econometrics literature. *arXiv preprint arXiv:2201.01194*. MR4602950
[25] Savarese, D. (2013). Common terminology criteria for adverse events. *UpToDate Waltham, MA: UpToDate* 1–9.
[26] Tchetgen Tchetgen, E. J., Robins, J. M. and Rotnitzky, A. (2010). On doubly robust estimation in a semiparametric odds ratio model. *Biometrika* **97** 171–180. MR2594425

[27] Tsiatis, A. A. (2006). *Semiparametric Theory and Missing Data*. Springer. MR2233926

[28] Uddin, M., Groenwold, R. H., Ali, M. S., de Boer, A., Roes, K. C., Chowdhury, M. A., Klungel, O. H. et al. (2016). Methods to control for unmeasured confounding in pharmacoepidemiology: an overview. *International Journal of Clinical Pharmacy* **38** 714–723.

[29] van der Laan, M. (2017). A generally efficient targeted minimum loss based estimator based on the highly adaptive lasso. *The International Journal of Biostatistics* **13**. MR3724476

[30] Van der Laan, M. J., Polley, E. C. and Hubbard, A. E. (2007). Super learner. *Statistical Applications in Genetics and Molecular Biology* **6**. MR2349918

[31] Van der Vaart, A. W. (2000). *Asymptotic Statistics* **3**. Cambridge University Press. MR1652247

[32] VanderWeele, T. J. and Knol, M. J. (2014). A tutorial on interaction. *Epidemiologic Methods* **3** 33–72.

[33] Vansteelandt, S. and Goetghebeur, E. (2003). Causal inference with generalized structural mean models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **65** 817–835. MR2017872

[34] Vo, T.-T., Ye, T., Ertefaie, A., Roy, S., Flory, J., Hennessy, S., Vansteelandt, S. and Small, D. S. (2024). Supplement for "Structural mean models for instrumented difference-in-differences". https://doi.org/10.1214/24-EJS2313SUPP.

[35] Wager, S. and Walther, G. (2015). Adaptive concentration of regression trees, with application to random forests. *arXiv preprint arXiv:1503.06388*.

[36] Wang, L. and 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. MR3798877

[37] Wing, C., Simon, K. and Bello-Gomez, R. A. (2018). Designing difference in difference studies: best practices for public health policy research. *Annu. Rev. Public Health* **39** 453–469.

[38] Ye, T., Ertefaie, A., Flory, J., Hennessy, S. and Small, D. S. (2022). Instrumented difference-in-differences. *Biometrics*. MR4606300

[39] Zhang, X., Faries, D. E., Li, H., Stamey, J. D. and Imbens, G. W. (2018). Addressing unmeasured confounding in comparative observational research. *Pharmacoepidemiology and Drug Safety* **27** 373–382.