Fuzzy Differences-in-Differences*

Clément de Chaisemartin† Xavier D'Haultfœuille‡

October 23, 2015

Abstract

In many applications of the differences-in-differences (DID) method, the treatment increases more in the treatment group, but some units are also treated in the control group. In such fuzzy designs, a popular estimator of treatment effects is the DID of the outcome divided by the DID of the treatment, or OLS and 2SLS regressions with time and group fixed effects estimating weighted averages of this ratio across groups. We start by showing that when the treatment also increases in the control group, this ratio estimates a causal effect only if treatment effects are homogenous in the two groups. Even when the distribution of treatment is stable, it requires that treatment effects be constant over time. As this assumption is not always applicable, we propose two alternative estimators. The first estimator relies on a generalization of common trends assumptions to fuzzy designs, while the second extends the changes-in-changes estimator of Athey & Imbens (2006). When the distribution of treatment changes in the control group, treatment effects are partially identified. Finally, we prove that our estimators are asymptotically normal and use them to revisit applied papers using fuzzy designs.

Keywords: differences-in-differences, changes-in-changes, quantile treatment effects, partial identification, returns to education.

JEL Codes: C21, C23

*This paper is a merged and revised version of de Chaisemartin & D'Haultfœuille (2014) and de Chaisemartin (2013). We thank Yannick Guyonvarch for excellent research assistance, and are very grateful to Esther Duflo and Erica Field for sharing their data with us. We also want to thank Alberto Abadie, Joshua Angrist, Stéphane Bonhomme, Marc Gurgand, Guido Imbens, Rafael Lalive, Thierry Magnac, Blaise Melly, Roland Rathelot, Bernard Salanié, Frank Vella, Fabian Waldinger, Yichong Zhang and participants at various conferences and seminars for their helpful comments.

†University of Warwick, clement.de-chaisemartin@warwick.ac.uk

‡CREST, xavier.dhaultfoeuille@ensae.fr

1 Introduction

Difference-in-differences (DID) is a popular method to evaluate the effect of a treatment in the absence of experimental data. In its basic version, a "control group" is untreated at two dates, whereas a "treatment group" becomes treated at the second date. If the effect of time is the same in both groups, the so-called common trends assumption, one can measure the effect of the treatment by comparing the evolution of the outcome in both groups. DID can be used with panel or repeated cross-section data, when a policy is implemented at a given date in some groups but not in others. It can also be used when a policy affects individuals born after a given date. In such instances, birth cohort plays the role of time.

However, in many applications of the DID method, the treatment rate or intensity increases more in some groups than in others, but there is no group which experiences a sharp change in treatment, and there is also no group which remains fully untreated. In such fuzzy designs, a popular estimator of treatment effects is the DID of the outcome divided by the DID of the treatment, an estimator referred to as the Wald-DID. For instance, Duflo (2001) uses a school construction program in Indonesia to measure returns to education. The author uses districts where many schools were constructed as a treatment group, and districts where few schools were constructed as a control group. Years of schooling for cohorts born after the program increased more in treatment districts. The author then estimates returns to schooling through a 2SLS regression in which dummies for cohorts benefiting from the program and for being born in treatment districts are used as controls, while the instrument is the interaction of these two dummies. The coefficient for treatment in this regression is the Wald-DID. A number of papers also estimate 2SLS regressions with time and group fixed effects and with a function of time and group as the excluded instrument, or OLS regressions at the group $\times$ period level with time and group fixed effects. In our supplementary material, we show that the coefficient of treatment in these two regressions is a weighted average of Wald-DIDs across groups. Such estimators have been frequently used by economic researchers. From 2010 to 2012, 10.1% of all papers published by the American Economic Review estimate either a simple Wald-DID, or the aforementioned IV or OLS regression. Excluding lab experiments and theory papers, this proportion raises to 19.7%.[1] Still, to our knowledge no paper has studied whether these estimators estimate a causal effect in models with heterogeneous treatment effects.

This papers makes the following contributions. We start by showing that the Wald-DID estimand is equal to a local average treatment effect (LATE) only if two strong assumptions are satisfied. First, time should have the same effect on all counterfactual outcomes, thus implying that the effect of the treatment should not vary over time. This assumption is often not applicable. For instance, in Duflo (2001) it requires that the wage gap between high school graduates born in younger and older cohorts should be the same had they not completed high

[1] Detailed results of our literature review can be found in de Chaisemartin & D'Haultfœuille (2015).

school. If they had not completed high school, graduates of every cohort would have entered the labor market earlier, and would have had more labor market experience by the time their wages are observed. As returns to experience tend to be concave (see Mincer & Jovanovic, 1979), the wage gap between graduates born in younger and older cohorts would presumably have been lower if they had not completed high school. Second, when treatment increases both in the treatment and in the control group, treatment effects should be homogenous in the two groups. Indeed, in such instances the Wald-DID is equal to a weighted difference between the LATE of treatment and control group units switching treatment over time. This weighted difference can be interpreted as a causal effect only if these two LATEs are equal. The weights received by each LATE can be estimated. In Duflo (2001), years of education increased substantially both in treatment and in control districts, so the Wald-DID in this paper is equal to a weighted difference between returns to schooling in treatment and control districts, and returns in the control group receive a large negative weight. This weighted difference estimates a causal effect only if returns to schooling are equal in the two groups of districts. This might be violated as control districts are more developed and could therefore have different returns. The IV and OLS regressions we study in our supplementary material suffer from the same problem. They both estimate a weighted sum of LATEs, with potentially many negative weights as we illustrate by estimating these weights in two applications.

Second, we propose two alternative estimators for the same LATE when the distribution of treatment is stable over time in the control group. Our first estimator, which we refer to as the time-corrected Wald ratio (Wald-TC), is a natural generalization of DID to fuzzy designs. It relies only on common trends assumptions between the treatment and the control group, within subgroups of units sharing the same treatment at the first date. Our second estimator, which we refer to as the changes-in-changes Wald ratio (Wald-CIC), generalizes the changes-in-changes estimator introduced by Athey & Imbens (2006) to fuzzy designs. It relies on the assumption that a control and a treatment unit with the same outcome and the same treatment at the first period will also have the same outcome at the second period.[2] Hereafter, we refer to this condition as the common changes assumption. Our Wald-TC and Wald-CIC estimators both have advantages and drawbacks, which we discuss later in the paper.

Third, we show that under the same common trends and common changes assumptions as those underlying the Wald-TC and Wald-CIC estimands, the same LATE can be bounded when the distribution of treatment changes over time in the control group. The smaller this change, the tighter the bounds. Fourth, we show how these results extend to settings with many group and periods, and how one can incorporate covariates in the analysis. Fifth, we consider estimators of the Wald-DID, Wald-TC, and Wald-CIC estimands, both with and without covariates. We show that they are asymptotically normal and prove the consistency

[2]Strictly speaking, the model in Athey & Imbens (2006) and our CIC model do not impose this restriction if one allows the unobserved determinant of the outcome to change over time. We still find this presentation of the CIC assumptions very helpful for pedagogical purposes.

of the bootstrap in some cases. Importantly, all our estimators allow for continuous covariates, and for some of them we show how to account for clustering.

Finally, we use our results to revisit findings in Duflo (2001) on returns to education. The distribution of schooling substantially changed in the control group used by the author, so using our Wald-CIC or Wald-TC estimators with her groups would only yield wide and uninformative bounds. Therefore, we use a different control group where the distribution of schooling did not change. Our Wald-DID estimate with these new groups is more than twice as large as the author's. The difference between these two estimates could stem from the fact that districts where years of schooling increased less also have higher returns to education. This would bias downward the estimate in Duflo (2001), while our estimator does not rely on any treatment effect homogeneity assumption. On the other hand, the validity of our Wald-DID still relies on the assumption that time has the same effect on all potential outcomes, which is not warranted in this context as we explained above. Because the Wald-TC and Wald-CIC do not rely on this assumption, we choose them as our favorite estimates. They both lie in between the two Wald-DIDs.

Overall, our paper shows that to do DID in fuzzy designs, researchers must find a control group in which treatment is stable over time to point identify treatment effects without having to assume that treatment effects are homogeneous. In such instances, three estimators are available: the standard Wald-DID estimator, and our Wald-TC and Wald-CIC estimators.[3] While the former estimator requires that treatment effects do not change over time, the latter estimators do not rely on this assumption. In practice, using one or the other estimator can make a substantial difference, as we show in our application.

Though to our knowledge, we are the first to study fuzzy DID estimators in models with heterogeneous treatment effects, our paper is related to several other papers in the DID and panel literature. Blundell et al. (2004) and Abadie (2005) consider a conditional version of the common trends assumption in sharp DID designs, and adjust for covariates using propensity score methods. Our Wald-DID estimator with covariates is related to their estimators. Bonhomme & Sauder (2011) consider a linear model allowing for heterogeneous effects of time, and show that in sharp designs it can be identified if the idiosyncratic shocks are independent of the treatment and of the individual effects. Our Wald-CIC estimator builds on Athey & Imbens (2006) and is also related to the estimator of D'Haultfœuille et al. (2013), who study the possibly nonlinear effects of a continuous treatment using repeated cross sections. Finally, Chernozhukov, Fernández-Val, Hahn & Newey (2013) consider a location-scale panel data model (see their Assumption 4). Their idea of using always and never treated units in the panel to recover the location and scale time effects is related to our idea of using groups where treatment is stable to recover time effects.[4] Our paper is also related to several papers in

[3] A stata package computing these estimators is available on the authors' webpages.

[4] There are also differences between our approaches. Their location and scale parameters do not depend on

the partial identification literature. In particular, our bounds are related to those in Manski (1990), Horowitz & Manski (1995), and Lee (2009).

The remainder of the paper is organized as follows. In Section 2 we introduce our framework. In Section 3 we present our identification results in a simple setting with two groups, two periods, a binary treatment, and no covariates. Section 4 considers extensions to settings with many periods and groups, covariates, or a non-binary treatment. Section 5 considers inference. In section 6 we revisit results from Duflo (2001). Section 7 concludes. The appendix gathers the main proofs. Due to a concern for brevity, some further results, our literature review, two supplementary applications, and additional proofs are deferred to our supplementary material (see de Chaisemartin & D'Haultfœuille, 2015).

2 Framework

We are interested in measuring the effect of a treatment D on some outcome. For now, we assume that treatment is binary. $Y(1)$ and $Y(0)$ denote the two potential outcomes of the same individual with and without treatment. The observed outcome is $Y = DY(1) + (1-D)Y(0)$.

We assume that the data at our disposal can be divided into "time periods" represented by a random variable T. If the analyst works with panel or repeated cross-sections data, time periods are dates. But in many DID papers, time periods are cohorts of the same population born in different years (see, e.g., Duflo, 2001). While with panel or repeated cross-sections data, each unit is or could be observed at both dates, with cohort data this is not the case. In what follows, we do not index observations by time, to ensure that our framework can apply to the three types of data. Referring to the panel data case is sometimes useful to convey the intuition of our results. However, our analysis is more targeted to the repeated cross-sections and cohort data cases: observing units at both dates open possibilities we do not explore here.

We also assume that the data can be divided into groups represented by a random variable G. In this section and in the next, we focus on the simplest possible case where there are only two groups, a "treatment" and a "control" group, and two periods of time. G is a dummy for units in the treatment group and T is a dummy for the second period. Contrary to the standard "sharp" DID setting where $D = G \times T$, we consider a "fuzzy" setting where $D \neq G \times T$. Some units may be treated in the control group or at period 0, and all units are not necessarily treated in the treatment group at period 1. However, we assume that the treatment rate increased more between period 0 and 1 in the treatment than in the control group.

We now introduce notations that we use throughout the paper. For any random variable R, let $\mathcal{S}(R)$ denote its support. Let also R_{gt} and R_{dgt} be two other random variables such that

the treatment while our Wald-TC (resp. Wald-CIC) estimator uses treatment specific additive shifts (resp. quantile-quantile transforms) to account for the effect of time; our Wald-TC estimator is not compatible with a location-scale model. Overall, our estimands are unrelated to theirs.

$R_{gt} \sim R|\, G = g, T = t$ and $R_{dgt} \sim R|\, D = d, G = g, T = t$, where $\sim$ denotes equality in distribution. Let F_R and $F_{R|S}$ denote the cumulative distribution function (cdfs) of R and its cdf conditional on S. For any event A, $F_{R|A}$ is the cdf of R conditional on A. With a slight abuse of notation, $P(A)F_{R|A}$ should be understood as 0 when $P(A) = 0$.

We consider the following model for the potential outcomes and the treatment:

$$\begin{aligned} Y(d) &= h_d(U_d, T), \qquad d \in \{0, 1\}, \\ D &= 1\{V \geq v_{GT}\}, \quad v_{G0} = v_{00} \text{ does not depend on G.} \end{aligned} \tag{1}$$

The model on potential outcomes is very general because at this stage, h_d is left unrestricted. We also impose a latent index model for the treatment (see, e.g., Vytlacil, 2002), where the threshold depends both on time and group. In such a model, V may be interpreted as the propensity to be treated. Because we do not impose any restriction on the distribution of V, the assumption that v_{G0} does not depend on G is just a normalization.

In addition to this model, we maintain the following assumptions throughout the paper.

Assumption 1 *(Time invariance within groups)*
For $d \in \mathcal{S}(D)$, $(U_d, V) \perp\!\!\!\perp T \,|\, G$.

Assumption 2 *(First stage)*
$E(D_{11}) > E(D_{10})$, and $E(D_{11}) - E(D_{10}) > E(D_{01}) - E(D_{00})$.

Assumption 1 requires that the joint distribution of unobserved variables be stable over time in each group. In other words, the composition of each group should not change over time. This assumption could be violated if there is endogenous "migration" from one group to another. However, DID identification strategies always rely on this assumption. Assumption 2 is just a way to define the treatment and the control group in our fuzzy setting. First, the treatment should increase in at least one group. If not, one can redefine the treatment variable as $\tilde{D} = 1 - D$. Then, the treatment group is the one experiencing the larger increase of its treatment rate.

Before turning to identification, it is useful to define four subpopulations of interest. The model 1 and Assumption 1 imply that $P(D_{gt} = 1) = P(V \geq v_{gt}|G = g)$. Therefore, Assumption 2 implies $v_{11} < v_{00}$. Let

$$\begin{aligned} AT &= \{V \geq v_{00}, G = 1\} \cup \{V \geq \max(v_{00}, v_{01}), G = 0\}, \\ NT &= \{V < v_{11}, G = 1\} \cup \{V < \min(v_{00}, v_{01}), G = 0\}, \\ S_1 &= \{V \in [v_{11}, v_{00}), G = 1\}, \\ S_0 &= \{V \in [\min(v_{00}, v_{01}), \max(v_{00}, v_{01})), G = 0\}. \end{aligned}$$

AT stands for "always treated", and refers to units with a taste for treatment above the threshold at both periods. NT stands for "never treated", and refers to units with a taste for

treatment below the threshold at both periods. S_1 stands for "treatment group switchers", and refers to treatment group units with a taste for treatment between the second and first period thresholds. S_0 stands for "control group switchers", and refers to control group units with a taste for treatment between the two thresholds.

When the treatment rate is stable in the control group, time affects selection into treatment only in the treatment group. Table 1 below considers an example. At both dates, untreated units in the control group belong to the NT subgroup, while treated units belong to the AT subgroup. On the other hand, untreated units in the treatment group in period 0 belong either to the NT or S_1 subgroup, while in period 1 they only belong to the NT subgroup. Conversely, treated units in period 0 only belong to the AT subgroup, while in period 1 they either belong to the NT or S_1 subgroup.

	Period 0	Period 1
Control Group	Always Treated: Y(1)	Always Treated: Y(1)
	Never Treated: Y(0)	Never Treated: Y(0)
Treatment Group	Always Treated: Y(1)	Always Treated: Y(1)
	Switchers: Y(0)	Switchers: Y(1)
	Never Treated: Y(0)	Never Treated: Y(0)

Table 1: Populations of interest when $P(D_{00} = 0) = P(D_{01} = 0)$.

On the other hand, when the treatment rate changes in the control group, time affects selection into treatment in both groups. Table 2 below considers an example where the treatment rate increases in the control group. Untreated units in the control group in period 0 belong either to the NT or S_0 subgroup, while in period 1 they only belong to the NT subgroup. Conversely, treated units in period 0 only belong to the AT subgroup, while in period 1 they either belong to the NT or S_0 subgroup.

	Period 0	Period 1
Control Group	Always Treated: Y(1)	Always Treated: Y(1)
	Switchers: Y(0)	Switchers: Y(1)
	Never Treated: Y(0)	Never Treated: Y(0)
Treatment Group	Always Treated: Y(1)	Always Treated: Y(1)
	Switchers: Y(0)	Switchers: Y(1)
	Never Treated: Y(0)	Never Treated: Y(0)

Table 2: Populations of interest when $P(D_{01} = 1) > P(D_{00} = 1)$.

Our identifications results focus on treatment group switchers. Our parameters of interest are their Local Average Treatment Effect (LATE) and Local Quantile Treatment Effects (LQTE), which are respectively defined by

$$\begin{aligned} \Delta &= E\left(Y_{11}(1) - Y_{11}(0)|S_1\right), \\ \tau_q &= F^{-1}_{Y_{11}(1)|S_1}(q) - F^{-1}_{Y_{11}(0)|S_1}(q), \quad q \in (0,1). \end{aligned}$$

We focus on this subpopulation because our assumptions either lead to point identification of Δ and τ_q, or at least to relatively tight bounds. On the other hand, our assumptions most often lead to wide and uninformative bounds for the average treatment effect and for quantile treatment effects.

3 Identification

3.1 Identification using a Wald-DID ratio

We first investigate the commonly used strategy of running an IV regression of the outcome on the treatment with time and group as included instruments, and the interaction of the two as the excluded instrument. The estimand arising from this regression is the Wald-DID defined by $W_{DID} = DID_Y/DID_D$ where, for any random variable R, we let

$$DID_R = E(R_{11}) - E(R_{10}) - \left(E(R_{01}) - E(R_{00})\right).$$

We consider a set of assumptions under which this estimand can receive a causal interpretation.

Assumption 3 *(Common trends)*

$E(h_0(U_0, 1) - h_0(U_0, 0)|G)$ *does not depend on* G.

Assumption 4 *(Common average effect of time on both potential outcomes)*

$E(h_1(U_1,1) - h_1(U_1,0)|G, V \geq v_{00}) = E(h_0(U_0,1) - h_0(U_0,0)|G, V \geq v_{00})$.

Assumption 3 requires that the mean of $Y(0)$ follow the same evolution over time in the treatment and control groups. This assumption is not specific to the fuzzy setting we are considering here: DID in sharp settings also rely on this assumption (see, e.g., Abadie, 2005). Assumption 4 requires that in both groups, the mean of $Y(1)$ and $Y(0)$ follow the same evolution over time among units treated in period 0. This is equivalent to assuming that the average treatment effect in this population does not change over time:

$$E(h_1(U_1,1) - h_0(U_0,1)|G, V \geq v_{00}) = E(h_1(U_1,0) - h_0(U_0,0)|G, V \geq v_{00}).$$

This assumption is specific to the fuzzy setting.

Theorem 3.1 *Assume that Model* (1) *and Assumptions 1-4 are satisfied. Let* $\alpha = \frac{P(D_{11}=1)-P(D_{10}=1)}{DID_D}$.

$$W_{DID} = \alpha E(Y_{11}(1) - Y_{11}(0)|S_1) + (1-\alpha)E(Y_{01}(1) - Y_{01}(0)|S_0).$$

When the treatment rate increases in the control group, $\alpha > 1$ so the Wald-DID is equal to a weighted difference of the LATEs of treatment and control group switchers in period 1. This can be seen from Table 2. In both groups, the evolution of the mean outcome between period 0 and 1 is the sum of three things: the effect of time on the mean of $Y(0)$ for never treated and switchers; the effect of time on the mean of $Y(1)$ for always treated; the average effect of the treatment for switchers. Under Assumptions 3 and 4, the effect of time in both groups cancel one another out. The Wald-DID is finally equal to the weighted difference between treatment and control group switchers' LATEs.

This weighted difference may not receive a causal interpretation. It might for instance be negative, while both $E(Y_{11}(1) - Y_{11}(0)|S_1)$ and $E(Y_{01}(1) - Y_{01}(0)|S_0)$ are positive. If one is ready to further assume that these two LATEs are equal, the Wald-DID is then equal to $E(Y_{11}(1) - Y_{11}(0)|S_1)$. But $E(Y_{11}(1) - Y_{11}(0)|S_1) = E(Y_{01}(1) - Y_{01}(0)|S_0)$ is a strong restriction on the heterogeneity of the treatment effect. To better understand why it is needed, let us consider a simple example in which all control group units have a treatment effect equal to +2, while all treatment group units have a treatment effect equal to +1. Let us also assume that time has no effect on the outcome, and that the treatment rate increases twice as much in the treatment than in the control group. Then, $W_{DID} = 2/3 \times 1 - 1/3 \times 2 = 0$: the lower increase of the treatment rate in the control group is exactly compensated by the fact that the treatment effect is higher in this group. The Wald-DID does not estimate the treatment effect in any of the two groups, or a weighted average of the two.

When the treatment rate diminishes in the control group, $\alpha < 1$ so the Wald-DID is equal to a weighted average of the LATEs of treatment and control group switchers in period 1. This

quantity satisfies the no sign-reversal property: if the treatment effect is of the same sign for everybody in the population, the Wald-DID is of that sign. Finally, when the treatment rate is stable over time in the control group, $\alpha = 1$ so the Wald-DID is equal to the LATE of treatment group switchers.

But even when the treatment rate is stable in the control group, the Wald-DID relies on the assumption that time has the same effect on both potential outcomes, at least among units treated in the first period. Under Assumptions 1-3 alone, one can show that W_{DID} is equal to the same quantity as in Theorem 3.1, plus a bias term equal to

$$\frac{1}{DID_D}\left[E(C_1 - C_0|V \geq v_{00}, G = 1)P(D_{10} = 1) - E(C_1 - C_0|V \geq v_{00}, G = 0)P(D_{00} = 1)\right],$$

where $C_d = h_d(U_d, 1) - h_d(U_d, 0)$. Assumption 5 ensures that this bias term is equal to 0. Otherwise, it might very well differ from 0.[5]

To understand why this restriction is needed, consider a simple example. First, assume that in period 0, $Y(1) = Y(0)$: treatment has no effect. Then, assume that time increases $Y(1)$ by 1 unit, while leaving $Y(0)$ unchanged. Finally, assume that the treatment rate went from to 20 to 50% in the treatment group, while it remained equal to 80% in the control group. Then, $DID_Y = 0.2 \times 1 + 0.3 \times 1 + 0.5 \times 0 - (0.8 \times 1 + 0.2 \times 0) = -0.3$. The first and third terms respectively come from the effect of time on the mean outcome of always and never treated in the treatment group. Similarly, the fourth and fifth terms respectively come from the effect of time on the mean outcome of always and never treated in the control group. Finally, the second term comes from the average treatment effect among treatment group switchers. Therefore, $W_{DID} = -1$, while every unit in the population has a treatment effect equal to 1 in period 1, and to 0 in period 0.

3.2 Identification using a time-corrected Wald ratio

In this section, we consider an alternative estimand to W_{DID}. Instead of relying on Assumptions 3 and 4, it relies on the following assumption:

Assumption 5 *(Common trends within treatment status at date 0)*

$E(h_0(U_0, 1) - h_0(U_0, 0)|G, V < v_{00})$ *and* $E(h_1(U_1, 1) - h_1(U_1, 0)|G, V \geq v_{00})$ *do not depend on* G.

Assumption 5 requires that the mean of $Y(0)$ (resp. $Y(1)$) follow the same evolution over time among treatment and control group units that were untreated (resp. treated) at period 0.

[5] Assuming that $E(C_0 - C_1|V < v_0, G)$ does not depend on G is not sufficient to ensure that the bias is equal to 0, unless $P(D_{00} = 1) = P(D_{10} = 1)$.

Let $\delta_d = E(Y_{d01}) - E(Y_{d00})$ denote the change in the mean outcome between period 0 and 1 for control group units with treatment status d. Then, let

$$W_{TC} = \frac{E(Y_{11}) - E(Y_{10} + \delta_{D_{10}})}{E(D_{11}) - E(D_{10})}.$$

W_{TC} stands for "time-corrected Wald". When the outcome is bounded, let $\underline{y}$ and $\overline{y}$ respectively denote the lower and upper bounds of its support. For any $g \in \mathcal{S}(G)$, let $\lambda_{gd} = P(D_{g1} = d)/P(D_{g0} = d)$ be the ratio of the shares of people receiving treatment d in period 1 and period 0 in group g. For instance, $\lambda_{00} > 1$ when the share of untreated observations increases in the control group between period 0 and 1. For any real number x, let $M_0(x) = \max(0, x)$ and $m_1(x) = \min(1, x)$. Let also, for $d \in \{0, 1\}$,

$$\begin{aligned}
\underline{F}_{d01}(y) &= M_0\left[1 - \lambda_{0d}(1 - F_{Y_{d01}}(y))\right] - M_0(1 - \lambda_{0d})\mathbb{1}\{y < \overline{y}\}, \\
\overline{F}_{d01}(y) &= m_1\left[\lambda_{0d} F_{Y_{d01}}(y)\right] + (1 - m_1(\lambda_{0d}))\mathbb{1}\{y \geq \underline{y}\}.
\end{aligned}$$

Then define $\underline{\delta}_d = \int y d\overline{F}_{d01}(y) - E(Y_{d00})$ and $\overline{\delta}_d = \int y d\underline{F}_{d01}(y) - E(Y_{d00})$ and let

$$\underline{W}_{TC} = \frac{E(Y_{11}) - E(Y_{10} + \overline{\delta}_{D_{10}})}{E(D_{11}) - E(D_{10})}, \; \overline{W}_{TC} = \frac{E(Y_{11}) - E(Y_{10} + \underline{\delta}_{D_{10}})}{E(D_{11}) - E(D_{10})}.$$

Theorem 3.2 *Assume that Model* (1) *and Assumptions 1-2 and 5 are satisfied.*

1. *If* $0 < P(D_{01} = 1) = P(D_{00} = 1) < 1$, $W_{TC} = \Delta$.
2. *If* $0 < P(D_{01} = 1) \neq P(D_{00} = 1) < 1$ *and* $P(\underline{y} \leq Y(d) \leq \overline{y}) = 1$ *for* $d \in \{0, 1\}$, $\underline{W}_{TC} \leq \Delta \leq \overline{W}_{TC}$.[6]

Note that

$$W_{TC} = \frac{E(Y|G = 1, T = 1) - E(Y + (1 - D)\delta_0 + D\delta_1|G = 1, T = 0)}{E(D|G = 1, T = 1) - E(D|G = 1, T = 0)}.$$

This is almost the Wald ratio with time as the instrument considered first by Heckman & Robb (1985), except that we have $Y + (1 - D)\delta_0 + D\delta_1$ instead of Y in the second term of the numerator. This difference arises because in our model time is not a standard instrument: it is directly included in the outcome equation. When the treatment rate is stable in the control group we can identify the direct effect of time on $Y(0)$ and $Y(1)$ by looking at how the mean outcome of untreated and treated units changes over time in this group. Under Assumption 5, this direct effect is the same in the two groups for units sharing the same treatment in the first period. As a result, we can add these changes to the outcome of untreated and treated units in the treatment group in period 0, to recover the mean outcome we would have observed in this group in period 1 if switchers had not changed their treatment between the two periods. This is what $(1 - D)\delta_0 + D\delta_1$ does. Therefore, the numerator of W_{TC} is equal to the effect

[6] It is not difficult to show that these bounds are sharp. We omit the proof due to a concern for brevity.

of time on the outcome that only goes through its effect on selection into treatment. Once properly normalized, this yields the LATE of treatment group switchers.

The Wald-TC estimand generalizes the DID estimand to fuzzy settings, by using treatment-specific additive shifts to account for the effect of time. In sharp settings, the DID estimand accounts for the effect of time on the outcome by adding the evolution of the mean outcome between period 0 and 1 in the control group to the period 0 outcome of treatment group units. In fuzzy settings, the Wald-TC estimand accounts for the effect of time on the outcome by adding the evolution of the mean outcome between period 0 and 1 among untreated (resp. treated) units in the control group to the period 0 outcome of untreated (resp. treated) units in the treatment group.

When the treatment rate changes in the control group, the evolution of the outcome in this group can stem both from the direct effect of time on the outcome, and from its effect on selection into treatment. For instance, and as can be seen from Table 2, when the treatment rate increases in the control group, the difference between $E(Y_{101})$ and $E(Y_{100})$ arises both from the effect of time on $Y(1)$, and from the fact the former expectation is for always treated and switchers while the later is only for always treated. Therefore, we can no longer identify the direct effect of time on the outcome. However, when the outcome has bounded support, this direct effect can be bounded, because we know the percentage of the control group switchers account for. As a result, the LATE of treatment group switchers can also be bounded. The smaller the change of the treatment rate over time in the control group, the tighter the bounds.

When the treatment rate does not change much in the control group, the difference between W_{TC} and Δ is likely to be small. For instance, when the treatment rate increases in the control group, it is easy to show that under the Assumptions of Theorem 3.2, W_{TC} is equal to Δ plus the following bias term:

$$\begin{aligned} &\frac{P(D_{10}=0)\left(1-\frac{P(D_{01}=0)}{P(D_{00}=0)}\right)\left(E(Y_{01}(0)|S_0)-E(Y_{01}(0)|NT)\right)}{P(D_{11}=1)-P(D_{10}=1)} \\ -\quad &\frac{P(D_{10}=1)\left(1-\frac{P(D_{00}=1)}{P(D_{01}=1)}\right)\left(E(Y_{01}(1)|S_0)-E(Y_{01}(1)|AT)\right)}{P(D_{11}=1)-P(D_{10}=1)}. \end{aligned} \tag{2}$$

This term cancels if $P(D_{01}=1)=P(D_{00}=1)$, but also if

$$U_0|S_0, G=0 \sim U_0|NT, G=0 \text{ and } U_1|S_0, G=0 \sim U_1|AT, G=0. \tag{3}$$

This assumption is not very appealing, as it requires that control group switchers have the same distribution of U_0 as never treated, and the same distribution of U_1 as always treated. But Equations (2) and (3) still show that when the treatment rate does not change much in the control group, W_{TC} is close to Δ unless switchers are extremely different from never and always treated.

Finally, note that when the treatment rate is stable in the control group, we have

$$W_{DID} = \frac{E(Y_{11}) - E(Y_{10} + \delta_{D_{00}})}{E(D_{11}) - E(D_{10})}.$$

When accounting for the effect of time on the outcome, W_{DID} weights δ_0 and δ_1 by $P(D_{00} = 0)$ and $P(D_{00} = 1)$, while W_{TC} weights these terms by $P(D_{10} = 0)$ and $P(D_{10} = 1)$. These two estimands are equal if and only if either $\delta_0 = \delta_1$ or $P(D_{00} = 1) = P(D_{10} = 1)$. Otherwise, they differ. The assumptions under which W_{DID} and W_{TC} rely are non-nested. W_{TC} requires more common trends assumptions between groups, but it does not require common trends assumptions between the two potential outcomes within groups. Therefore, testing $W_{DID} = W_{TC}$ is a joint test of Assumptions 1 and 3-5.

3.3 Identification using instrumented changes-in-changes

In this section, we consider a second alternative estimand to W_{DID} for continuous outcomes. This estimand is inspired from the CIC model in Athey & Imbens (2006). It crucially relies on a monotonicity assumption.

Assumption 6 *(Monotonicity)*

$U_d \in \mathbb{R}$ and $h_d(u,t)$ is strictly increasing in u for all $(d,t) \in \mathcal{S}(D) \times \mathcal{S}(T)$.

Assumption 6 requires that at each period, potential outcomes are strictly increasing functions of a scalar unobserved heterogeneity term. Hereafter, we refer to Assumptions 1-2 and 6 as to the IV-CIC model. The IV-CIC model generalizes the CIC model to fuzzy settings. Assumption 1 implies $U_d \perp\!\!\!\perp T \,|\, G$ and $V \perp\!\!\!\perp T \,|\, G$, which correspond to the time invariance assumption in Athey & Imbens (2006). As a result, the IV-CIC model imposes a standard CIC model both on Y and D. But Assumption 1 also implies $U_d \perp\!\!\!\perp T \,|\, G, V$: in each group, the distribution of, say, ability among people with a given taste for treatment should not change over time. Our results rely on this supplementary restriction.

The assumptions of the IV-CIC model have advantages and drawbacks with respect to those underlying the Wald-DID and Wald-TC estimands. For instance, one implication of Assumptions 1 and 5 is that the difference between the mean outcome of always treated in the treatment and in the control group should remain stable over time. This condition is not invariant to the scaling of the outcome, but it only restricts its first moment. On the other hand, the corresponding implication of Assumptions 1 and 6 is that the proportion of units in the treatment group among any quantile group of the always treated remains constant over time. For instance, if in period 0 70% of units in the first decile of always treated belonged to the treatment group, in period 1 there should still be 70% of treatment group units in the first decile.[7] This condition is invariant to the scaling of the outcome, but it restricts its entire

[7] Unfortunately, this condition is not testable as always treated are not observed.

distribution. When the treatment and the control groups have different outcome distributions in the first period (see e.g. Baten et al., 2014), the scaling of the outcome might have a large effect on the results, so using a model invariant to this scaling might be preferable. On the other hand, when the outcome distributions in the treatment and in the control group are similar in the first period, using a model that only restricts the first moment of the outcome might be preferable.

We also impose the assumption below, which is testable in the data.

Assumption 7 *(Data restrictions)*

1. $\mathcal{S}(Y_{dgt}) = \mathcal{S}(Y) = [\underline{y}, \overline{y}]$ *with* $-\infty \leq \underline{y} < \overline{y} \leq +\infty$, *for* $(d, g, t) \in \mathcal{S}(D) \times \mathcal{S}(G) \times \mathcal{S}(T)$.
2. $F_{Y_{dgt}}$ *is continuous on* $\mathbb{R}$ *and strictly increasing on* $\mathcal{S}(Y)$, *for* $(d, g, t) \in \mathcal{S}(D) \times \mathcal{S}(G) \times \mathcal{S}(T)$.

The first condition requires that the outcome have the same support in each of the eight treatment $\times$ group $\times$ period cell. This condition does not restrict the support to be bounded: $\underline{y}$ and $\overline{y}$ can be equal to $-$ and $+\infty$. Athey & Imbens (2006) make a similar assumption. Common support conditions might not be satisfied when outcome distributions differ in the treatment and in the control group, the very situations where CIC might be more appealing than DID. Athey & Imbens (2006) show that in such instances, quantile treatment effects are still point identified over a large set of quantiles, while the average treatment effect can be bounded. Even though we do not present them here, similar results apply in fuzzy settings.

The second condition is satisfied if the distribution of Y is continuous with positive density in each of the eight groups $\times$ periods $\times$ treatment status cells. With a discrete outcome, Athey & Imbens (2006) show that one can bound treatment effects under their assumptions. Similar results apply in fuzzy settings, but as CIC bounds for discrete outcomes are often not very informative, we do not present them here.

Let $Q_d(y) = F^{-1}_{Y_{d01}} \circ F_{Y_{d00}}(y)$ be the quantile-quantile transform of Y from period 0 to 1 in the control group conditional on $D = d$. This transform maps y at rank q in period 0 into the corresponding y' at rank q in period 1. Let also $H_d(q) = F_{Y_{d10}} \circ F^{-1}_{Y_{d00}}(q)$ be the inverse quantile-quantile transform of Y from the control to the treatment group in period 0 conditional on $D = d$. This transform maps rank q in the control group into the corresponding rank q' in the treatment group with the same value of y. Finally, for any increasing function F on the real line, we denote by F^{-1} its generalized inverse:

$$F^{-1}(q) = \inf \{x \in \mathbb{R} : F(x) \geq q\}.$$

In particular, F^{-1}_R is the quantile function of the random variable R. We adopt the convention that $F^{-1}_R(q) = \inf \mathcal{S}(R)$ for $q < 0$, and $F^{-1}_R(q) = \sup \mathcal{S}(R)$ for $q > 1$.

Our identification results rely on the following lemma.

Lemma 3.1 *If Assumptions 1-2 and 6-7 hold and if* $P(D_{00}=d)>0$,

$$F_{Y_{11}(d)|S_1}(y)=\frac{P(D_{10}=d)H_d\circ(\lambda_{0d}F_{Y_{d01}}(y)+(1-\lambda_{0d})F_{Y_{01}(d)|S_0}(y))-P(D_{11}=d)F_{Y_{d11}}(y)}{P(D_{10}=d)-P(D_{11}=d)}.$$

This lemma shows that under our IV-CIC assumptions, $F_{Y_{11}(d)|S_1}$ is point identified when the treatment rate remains constant in the control group, as in this case $\lambda_{0d}=1$. Let

$$F_{CIC,d}(y)=\frac{P(D_{10}=d)H_d\circ(F_{Y_{d01}}(y))-P(D_{11}=d)F_{Y_{d11}}(y)}{P(D_{10}=d)-P(D_{11}=d)},$$
$$W_{CIC}=\frac{E(Y_{11})-E(Q_{D_{10}}(Y_{10}))}{E(D_{11})-E(D_{10})}.$$

When the treatment rate changes in the control group, $F_{Y_{11}(d)|S_1}$ is partially identified. Sharp bounds can be obtained using Lemma 3.1. For any cdf T_d, let

$$G_d(T_d)=\lambda_{0d}F_{Y_{d01}}+(1-\lambda_{0d})T_d,$$
$$C_d(T_d)=\frac{P(D_{10}=d)H_d\circ G_d(T_d)-P(D_{11}=d)F_{Y_{d11}}}{P(D_{10}=d)-P(D_{11}=d)}.$$

It follows from Lemma 3.1 that $C_d(F_{Y_{01}(d)|S_0})=F_{Y_{11}(d)|S_1}$. Moreover, one can show that $G_0(F_{Y_{01}(0)|S_0})=F_{Y_{01}(0)|V<v_{00}}$ and $G_1(F_{Y_{01}(1)|S_0})=F_{Y_{01}(1)|V\geq v_{00}}$. Therefore, the sharp lower bound on $F_{Y_{11}(d)|S_1}$ is

$$\min_{T_d\in\mathcal{D}} C_d(T_d) \quad \text{s.t. } (T_d,G_d(T_d),C_d(T_d))\in\mathcal{D}^3,$$

where $\mathcal{D}$ is the set of cdfs on $\mathcal{S}(Y)$.

It is difficult to derive a closed-form expression for the solution of this problem, because it corresponds to an infinite dimensional optimization problem with an infinite number of inequality constraints. We therefore consider simpler bounds, which are sharp under a simple testable assumption. Specifically, let $M_{01}(x)=\min(1,\max(0,x))$, and let

$$\underline{T}_d=M_{01}\left(\frac{\lambda_{0d}F_{Y_{d01}}-H_d^{-1}(\lambda_{1d}F_{Y_{d11}})}{\lambda_{0d}-1}\right),\ \overline{T}_d=M_{01}\left(\frac{\lambda_{0d}F_{Y_{d01}}-H_d^{-1}(\lambda_{1d}F_{Y_{d11}}+(1-\lambda_{1d}))}{\lambda_{0d}-1}\right),$$
$$\underline{F}_{CIC,d}(y)=\sup_{y'\leq y} C_d\left(\underline{T}_d\right)(y'),\ \overline{F}_{CIC,d}(y)=\inf_{y'\geq y} C_d\left(\overline{T}_d\right)(y'),$$
$$\underline{W}_{CIC}=\int y d\overline{F}_{CIC,1}(y)-\int y d\underline{F}_{CIC,0}(y),\ \overline{W}_{CIC}=\int y d\underline{F}_{CIC,1}(y)-\int y d\overline{F}_{CIC,0}(y),$$
$$\underline{\tau}_q=\max(\overline{F}^{-1}_{CIC,1}(q),\underline{y})-\min(\underline{F}^{-1}_{CIC,0}(q),\overline{y}),\ \overline{\tau}_q=\min(\underline{F}^{-1}_{CIC,1}(q),\overline{y})-\max(\overline{F}^{-1}_{CIC,0}(q),\underline{y}).$$

Finally, we introduce the two following conditions.

Assumption 8 *(Existence of moments)*

$\int|y|d\overline{F}_{CIC,d}(y)<+\infty$ *and* $\int|y|d\underline{F}_{CIC,d}(y)<+\infty$ *for* $d\in\{0,1\}$.

Assumption 9 *(Increasing bounds)*

For $(d,g,t) \in \mathcal{S}(D) \times \{0,1\}^2$, $F_{Y_{dgt}}$ is continuously differentiable, with positive derivative on the interior of $\mathcal{S}(Y)$. Moreover, $\underline{T}_d, \overline{T}_d, G_d(\underline{T}_d), G_d(\overline{T}_d), C_d(\underline{T}_d)$ and $C_d(\overline{T}_d)$ are increasing on $\mathcal{S}(Y)$.

Theorem 3.3 *Assume that Model (1) and Assumptions 1-2 and 6-7 hold.*

1. *If $0 < P(D_{01} = 1) = P(D_{00} = 1) < 1$, then $F_{CIC,d}(y) = F_{Y_{11}(d)|S_1}(y)$ for $d \in \{0,1\}$, $W_{CIC} = \Delta$ and $F^{-1}_{CIC,1}(q) - F^{-1}_{CIC,0}(q) = \tau_q$.*

2. *If $0 < P(D_{01} = 1) \neq P(D_{00} = 1) < 1$ and Assumption 8 is satisfied, then $F_{Y_{11}(d)|S_1}(y) \in [\underline{F}_{CIC,d}(y), \overline{F}_{CIC,d}(y)]$ for $d \in \{0,1\}$, $\Delta \in [\underline{W}_{CIC}, \overline{W}_{CIC}]$ and $\tau_q \in [\underline{\tau}_q, \overline{\tau}_q]$. Moreover, if Assumption 9 holds, these bounds are sharp.*

Our point identification results combine ideas from Imbens & Rubin (1997) and Athey & Imbens (2006). We seek to recover the distribution of, say, $Y(1)$ among switchers in the treatment $\times$ period 1 cell. On that purpose, we start from the distribution of Y among all treated observations of this cell. As shown in Table 1, those include both switchers and always treated. Consequently, we must "withdraw" from this distribution that of $Y(1)$ among always treated, exactly as in Imbens & Rubin (1997). But this last distribution is not observed. To reconstruct it, we adapt the ideas in Athey & Imbens (2006) and apply the quantile-quantile transform from period 0 to 1 among treated observations in the control group to the distribution of $Y(1)$ among treated units in the treatment group in period 0.

Intuitively, the quantile-quantile transform uses a double-matching to reconstruct the unobserved distribution. Consider an always treated in the treatment $\times$ period 0 cell. She is first matched to an always treated in the control $\times$ period 0 cell with same y. Those two always treated are observed at the same period of time and are both treated. Therefore, under Assumption 6 they must have the same u_1. Second, the control $\times$ period 0 always treated is matched to her rank counterpart among always treated of the control $\times$ period 1 cell. We denote y^* the outcome of this last observation. Because $U_1 \perp\!\!\!\perp T|G, V \geq v_{00}$, those two observations must also have the same u_1. Consequently, $y^* = h_1(u_1, 1)$, which means that y^* is the outcome that the treatment $\times$ period 0 cell unit would have obtained in period 1.

Note that

$$W_{CIC} = \frac{E(Y|G=1,T=1) - E((1-D)Q_0(Y) + DQ_1(Y)|G=1,T=0)}{E(D|G=1,T=1) - E(D|G=1,T=0)}.$$

Here again, W_{CIC} is almost the standard Wald ratio in the treatment group with T as the instrument, except that we have $(1-D)Q_0(Y) + DQ_1(Y)$ instead of Y in the second term of the numerator. $(1-D)Q_0(Y) + DQ_1(Y)$ accounts for the fact time has a direct effect on the outcome. When the treatment rate is stable in the control group, we can identify this

direct effect by looking at how the distribution of the outcome evolves in this group. We can then net out this direct effect in the treatment group. This is what $(1-D)Q_0(Y)+DQ_1(Y)$ does. Both W_{CIC} and W_{TC} proceed from the same logic, except that W_{TC} corrects for the effect of time through additive shifts, while W_{CIC} does so in a non-linear fashion. If $h_d(U_d,T) = a_d(U_d) + b_d(T)$ with $a_d(.)$ strictly increasing, Assumptions 5 and 6 are both satisfied. We then have $W_{CIC} = W_{TC}$.

Our partial identification results are obtained as follows. When $0 < P(D_{00}=1) \neq P(D_{01}=1) < 1$, the second matching described above collapses, because treated (resp. untreated) observations in the control group are no longer comparable in period 0 and 1. For instance, when the treatment rate increases in the control group, treated observations in the control group include only always treated in period 0. In period 1 they also include switchers, as is shown in Table 2. Therefore, we cannot match period 0 and period 1 observations on their rank anymore. However, under Assumption 1 the respective weights of switchers and always treated in period 1 are known. We can therefore derive best and worst case bounds for the distribution of the outcome for always treated in period 1, and match period 0 observations to their best and worst case rank counterparts.

If the support of the outcome is unbounded, $\underline{F}_{CIC,0}$ and $\overline{F}_{CIC,0}$ are proper cdf when $\lambda_{00} > 1$, but they are defective when $\lambda_{00} < 1$. When $\lambda_{00} < 1$, switchers belong to the group of treated observations in the control $\times$ period 1 cell (cf. Table 2). Their $Y(0)$ is not observed in period 1, so the data does not impose any restriction on $F_{Y_{01}(0)|S_0}$: it could be equal to 0 or to 1, hence the defective bounds. On the contrary, when $\lambda_{00} > 1$, switchers belong to the group of untreated observations in the control $\times$ period 1 cell, and under Assumption 1 we know that they account for $100(1-1/\lambda_{00})\%$ of this group. Consequently, we can use trimming bounds for $F_{Y_{01}(0)|S_0}$ (see Horowitz & Manski, 1995), hence the non-defective bounds. On the contrary, $\underline{F}_{CIC,1}$ and $\overline{F}_{CIC,1}$ are always proper cdf, while we could have expected them to be defective when $\lambda_{00} > 1$. This asymmetry stems from the fact that when $\lambda_{00} > 1$, setting $F_{Y_{01}(1)|S_0}(y) = 0$ would yield $F_{Y_{01}(1)|S_1}(y) > 1$ for values of y approaching $\overline{y}$, while setting $F_{Y_{01}(1)|S_0}(y) = 1$ would yield $F_{Y_{01}(1)|S_1}(y) < 0$ for values of y approaching $\underline{y}$.

The previous discussion implies that when $\mathcal{S}(Y)$ is unbounded and $\lambda_{00} < 1$, our bounds on Δ are infinite because our bounds for the cdf of $Y(0)$ of switchers are defective. Our bounds on τ_q are also infinite for low and high values of q. On the contrary, when $\lambda_{00} > 1$ our bounds on τ_q are finite for every $q \in (0,1)$. Our bounds on Δ are also finite provided $\underline{F}_{CIC,0}$ and $\overline{F}_{CIC,0}$ admit an expectation.

Finally, when the treatment rate changes in the control group, one can recover point identification if one is ready to impose the same supplementary assumption as in Equation (3).

3.4 Identification with a fully treated or fully untreated control group

Up to now, we have considered general fuzzy situations where the $P(D_{gt} = d)$ were restricted only by Assumption 2. An interesting special case, which is close to the sharp design, is when $P(D_{00} = 1) = P(D_{01} = 1) = P(D_{10} = 1) = 0$. In such instances, identification of the average treatment effect on the treated can be obtained under the same assumptions as those of the standard DID or CIC models.

Theorem 3.4 *Suppose that* $P(D_{00} = 1) = P(D_{01} = 1) = P(D_{10} = 1) = 0 < P(D_{11} = 1)$, $U_0 \perp\!\!\!\perp T|G$, *and the outcome equation of Model* (1) *is satisfied.*

1. *If Assumption 3 holds, then* $W_{DID} = W_{TC} = E(Y_{11}(1) - Y_{11}(0)|D = 1)$.
2. *If Assumptions 6 and 7 hold, then* $W_{CIC} = E(Y_{11}(1) - Y_{11}(0)|D = 1)$.

Hence, results of the sharp case extend to this intermediate case. Note that under Model (1) and Assumption 1, the treated population corresponds to S_1, so $E(Y_{11}(1)-Y_{11}(0)|D = 1) = \Delta$ under these additional assumptions.

Another special case of interest is when $P(D_{00} = 0) = P(D_{01} = 0) \in \{0, 1\}$. Such situations arise when a policy is extended to a previously a group, or when a program or a technology previously available in some geographic areas is extended to others (see our second supplementary application in de Chaisemartin & D'Haultfœuille (2015)). Theorem 3.1 applies in this special case, but not Theorems 3.2-3.3, as they require that $0 < P(D_{00} = 0) = P(D_{01} = 0) < 1$. In such instances, identification must rely on the assumption that time has the same effect on both potential outcomes. For instance, if $P(D_{00} = 1) = P(D_{01} = 1) = 1$ and $P(D_{10} = 1) < 1$, there are no untreated units in the control group that we can use to infer trends for untreated units in the treatment group. We must therefore use treated units, under the assumption that time has the same effect on both potential outcomes. Instead of the Wald-TC estimand, one could then use $\frac{E(Y_{11})-E(Y_{10}+\delta_1)}{E(D_{11})-E(D_{10})}$. Because $P(D_{00} = 1) = P(D_{01} = 1) = 1$, this actually amounts to using W_{DID}. We can also adapt our Wald-CIC estimand by considering the following assumption.

Assumption 10 *(Common effect of time on both potential outcomes)*

$h_0(h_0^{-1}(y, 1), 0) = h_1(h_1^{-1}(y, 1), 0)$ *for every* $y \in \mathcal{S}(Y)$.

Assumption 10 requires that time have the same effect on both potential outcomes: once combined with Equation (1) and Assumption 6, Assumption 10 implies that a treated and an untreated unit with the same outcome in period 0 also have the same outcome in period 1. This restriction is not implied by the IV-CIC assumptions we introduced in Section 3.3: Equation (1) and Assumption 6 alone only imply that two treated (resp. untreated) units with the same outcome in period 0 also have the same outcome in period 1. An example of

a structural function satisfying Assumption 10 is $h_d(U_d, T) = f(g_d(U_d), T)$ with $f(., t)$ and $g_d(.)$ strictly increasing. This shows that Assumption 10 does not restrict the effects of time and treatment to be homogeneous. Finally, Assumptions 4 and 10 are related, but they also differ on some respects. Assumption 4 restricts time to have the same average effect on the potential outcomes of always treated. Assumption 10 restricts time to have the same effect on the potential outcomes of units satisfying $Y(0) = Y(1)$ at the first period.

Under Assumption 10, if $P(D_{00} = d) = P(D_{01} = d) = 1$ we can use changes in the distribution of $Y(d)$ in the control group over time to identify the effect of time on $Y(1-d)$, hence allowing us to recover both $F_{Y_{11}(d)|S_1}$ and $F_{Y_{11}(1-d)|S_1}$.

Theorem 3.5 *If Assumptions 1-2, 6-7, and 10 hold, and $P(D_{00} = d) = P(D_{01} = d) = 1$ for some $d \in \{0, 1\}$,*

$$\frac{P(D_{10} = d)F_{Q_d(Y_{d10})}(y) - P(D_{11} = d)F_{Y_{d11}}(y)}{P(D_{10} = d) - P(D_{11} = d)} = F_{Y_{11}(d)|S_1}(y),$$
$$\frac{P(D_{10} = 1-d)F_{Q_d(Y_{1-d10})}(y) - P(D_{11} = 1-d)F_{Y_{1-d11}}(y)}{P(D_{10} = 1-d) - P(D_{11} = 1-d)} = F_{Y_{11}(1-d)|S_1}(y),$$
$$\frac{E(Y_{11}) - E(Q_d(Y_{10}))}{E(D_{11}) - E(D_{10})} = \Delta.$$

The estimands introduced in this theorem are very similar to those considered in the first point of Theorem 3.3, except that they apply the same quantile-quantile transform to all treatment units in period 0, instead of applying different transforms to units with a different treatment.

Finally, when $0 < P(D_{00} = 1) = P(D_{01} = 1) < 1$, Assumption 10 is testable. If it is satisfied, the quantile-quantile transforms Q_0 and Q_1 must be equal. When this test is not rejected, applying a weighted average of these two transforms to all treatment group units in period 0 might result in efficiency gains with respect to our Wald-CIC estimator.[8]

3.5 Panel data models

Model (1) is well suited for repeated cross sections or cohort data where we observe units only once. On the other hand, it implies a strong restriction on selection into treatment when panel data are available. As V does not depend on time, our selection equation implies that within each group, time can affect individuals' treatment decision in only one direction. Actually, all our results remain valid if U_d and V are indexed by time, provided that we rewrite Assumption 1 as follows: for $d \in \mathcal{S}(D)$, the distribution of $(U_{dt}, V_t)\,|G$ does not depend on t. Within each group, time could then induce some units to go from non-treatment to treatment, while having the opposite effect on other units.

We now discuss whether the common trends and monotonicity assumptions we introduced above are satisfied in standard panel data models. We index random variables by i, to distinguish individual effects from constant terms.

[8] We would like to thank an anonymous referee for pointing this out to us.

First, we consider the following model:

$$Y_{it} = \gamma_t + \alpha_i + \beta_i D_{it} + \varepsilon_{it}, \tag{4}$$

$$D_{it} = 1\{V_{it} \geq v_{G_i t}\}, \tag{5}$$

$$(\varepsilon_{i1}, V_{i1}, \alpha_i, \beta_i)|G_i \sim (\varepsilon_{i0}, V_{i0}, \alpha_i, \beta_i)|G_i. \tag{6}$$

The outcome equation has time and individual effects. It allows for heterogeneous but time invariant treatment effects which can be arbitrarily correlated with the treatment, the individual effect α_i, and the idiosyncratic shocks. Equation (6) requires that the distribution of $(\varepsilon_{it}, V_{it}, \alpha_i, \beta_i)|G_i$ does not depend on time. On the other hand, it does not restrict the cross-sectional dependence between ε_{it} and V_{it}, nor the serial dependence between $(\varepsilon_{i0}, V_{i0})$ and $(\varepsilon_{i1}, V_{i1})$. This implies in particular that in the first-difference equation, $D_{i1} - D_{i0}$ is endogenous in general. The Wald-DID estimand then amounts to instrumenting $D_{i1} - D_{i0}$ by G_i in this first-difference equation. It is easy to see that if Equations (4)-(6) hold, then Assumptions 1-6 are satisfied:[9] the additive separability of the time effect ensures that Assumptions 3, 5, and 6 are satisfied, while the time invariant treatment effects ensure that Assumption 4 is satisfied.

Second, we consider the following outcome equation instead of Equation (4):

$$Y_{it} = \gamma_t + \lambda_t D_{it} + \alpha_i + \beta_i D_{it} + \varepsilon_{it}. \tag{7}$$

Under Equation (7), Assumption 4 is no longer satisfied because treatment effects change over time. On the other hand, the effect of time is still additively separable from treatment and from the unobserved heterogeneity terms, so Equations (7) and (5)-(6) guarantee that Assumptions 1-2 and 5-6 are satisfied.

Then, we consider the following outcome equation:

$$Y_{it} = \gamma_t + \lambda_t D_{it} + \mu_t(\alpha_i + \beta_i D_{it} + \varepsilon_{it}). \tag{8}$$

Under Equation (8), Assumption 5 is no longer satisfied because time has an heterogeneous effect on the outcome. On the other hand, if Equations (8) and (5)-(6) hold, then Assumptions 1-2 and 6 are satisfied. To see this, define $h_d(u,t) = \gamma_t + \lambda_t d + \mu_t u$ and $U_{dit} = \alpha_i + \beta_i d + \varepsilon_{it}$.

Finally, we consider a last outcome equation:

$$Y_{it} = \gamma_t + \lambda_t D_{it} + \alpha_i + \beta_i D_{it} + \mu_t \varepsilon_{it}. \tag{9}$$

All our assumptions fail to hold under this fixed effects model with time-varying effects of the idiosyncratic shock. As above, Assumption 5 fails because time has heterogeneous effects

[9] As mentioned above, U_d and V should be indexed by time, and Assumption 1 should be rewritten as follows: for $d \in \mathcal{S}(D)$, the distribution of $(U_{dt}, V_t)\,|G$ is independent of t.

on the outcome. Assumption 6 also fails because the outcome can no longer be written as a function of time and a scalar unobserved term. Bonhomme & Sauder (2011) study a similar model with fixed effects and non-stationary idiosyncratic shocks. In the sharp case, they show that average and quantile treatment effects are identified if the idiosyncratic shocks are independent of treatment and of the fixed effects.

4 Extensions

In this section, we extend our analysis to situations where the data can be divided into several groups and several periods, where covariates are available, or where the treatment is non-binary. To generalize our results, we have to modify some of the assumptions we introduced above. To ease the comparison, we label these assumptions using suffixes. For instance Assumption 1X is similar to Assumption 1 except that it accounts for covariates X.

4.1 Multiple groups and time periods

Let us consider the case where the data can be divided into more than two groups and time periods. Let $G \in \{0, 1, ..., \overline{g}\}$ be the group a unit belongs to. Let $T \in \{0, 1, ..., \overline{t}\}$ be the period when she is observed. For any $(g, t) \in \mathcal{S}(G) \times \{1, ..., \overline{t}\}$, let $S_{gt} = \{V \in [\min(v_{gt-1}, v_{gt}), \max(v_{gt-1}, v_{gt})), G = g\}$ be the subset of group g which switches treatment status between $t-1$ and t. Also, let $S_t = \cup_{g=0}^{\overline{g}} S_{gt}$ denote the units switching between $t-1$ and t. Finally, let $S = \bigcup_{t=1}^{\overline{t}} S_t$ be the union of all switchers. At each date, we can partition the groups into three subsets, depending on whether their treatment rate is stable, increases, or decreases between $t-1$ and t. For every $t \in \{1, ..., \overline{t}\}$, let

$$\begin{aligned}
\mathcal{G}_{st} &= \{g \in \mathcal{S}(G) : E(D_{gt}) = E(D_{gt-1})\} \\
\mathcal{G}_{it} &= \{g \in \mathcal{S}(G) : E(D_{gt}) > E(D_{gt-1})\} \\
\mathcal{G}_{dt} &= \{g \in \mathcal{S}(G) : E(D_{gt}) < E(D_{gt-1})\},
\end{aligned}$$

and let $G_t^* = 1\{G \in \mathcal{G}_{it}\} - 1\{G \in \mathcal{G}_{dt}\}$. We introduce the following assumptions, which generalize Assumptions 3-5 to settings with multiple groups and periods (Assumptions 6 and 7 apply to this case without modifications).

Assumption 3*M* *(Common trends)*

For every $t \in \{1, ..., \overline{t}\}$, $E(h_0(U_0, t) - h_0(U_0, t-1)|G)$ *does not depend on* G.

Assumption 4*M* *(Common average effect of time on both potential outcomes)*

For every $t \in \{1, ..., \overline{t}\}$, $E(h_1(U_1, t) - h_1(U_1, t-1)|G, V \geq v_{Gt-1}) = E(h_0(U_0, t) - h_0(U_0, t-1)|G, V \geq v_{Gt-1})$.

Assumption 5*M* *(Common trends within treatment at previous period)*

For every $t \in \{1,...,\bar{t}\}$, $E(h_0(U_0,t)-h_0(U_0,t-1)|G,V<v_{Gt-1})$ *and* $E(h_1(U_1,t)-h_1(U_1,t-1)|G,V\geq v_{Gt-1})$ *do not depend on* G.

Theorem 4.1 below shows that when there is at least one group in which the treatment rate is stable between each pair of consecutive dates, combinations of these assumptions allow us to point identify Δ_w, a weighted average of LATEs over different periods:

$$\Delta_w = \sum_{t=1}^{\bar{t}} \frac{P(S_t)}{\sum_{t=1}^{\bar{t}} P(S_t)} E(Y(1)-Y(0)|S_t, T=t).$$

We also consider the following assumption, under which Δ_w is equal to the LATE among the whole population of switchers S.

Assumption 11 *(Monotonic evolution of treatment, and homogenous effects over time)*

1. *For every* $t \neq t' \in \{1,...,\bar{t}\}^2$ $\mathcal{G}_{it} \cap \mathcal{G}_{dt'} = \emptyset$.
2. *For every* $(t,t') \in \{1,...,\bar{t}\}^2$, $E(Y(1)-Y(0)|S_t, T=t') = E(Y(1)-Y(0)|S_t, T=1)$.

The first point of Assumption 11 requires that in every group, the treatment rate follows a monotonic evolution over time. The second point requires that switchers' LATE be constant over time.

For any random variable R and for any $g \neq g' \in \{-1,0,1\}^2$ and $t \in \{1,...,\bar{t}\}$ let

$$\begin{aligned}
DID^*_R(g,g',t) &= E(R|G^*_t=g, T=t) - E(R|G^*_t=g, T=t-1) \\
&- (E(R|G^*_t=g', T=t) - E(R|G^*_t=g', T=t-1)) \\
W^*_{DID}(g,g',t) &= \frac{DID^*_Y(g,g',t)}{DID^*_D(g,g',t)} \\
w_t &= \frac{DID^*_D(1,0,t)P(G^*_t=1) + DID^*_D(0,-1,t)P(G^*_t=-1)}{\sum_{t=1}^{\bar{t}} DID^*_D(1,0,t)P(G^*_t=1) + DID^*_D(0,-1,t)P(G^*_t=-1)} \\
w_{10|t} &= \frac{DID^*_D(1,0,t)P(G^*_t=1)}{DID^*_D(1,0,t)P(G^*_t=1) + DID^*_D(0,-1,t)P(G^*_t=-1)}.
\end{aligned}$$

Let also

$$\begin{aligned}
&\delta^*_{dt} = E(Y|D=d, G^*_t=0, T=t) - E(Y|D=d, G^*_t=0, T=t-1) \text{ for } d \in \{0,1\} \\
&W^*_{TC}(1,0,t) = \frac{E(Y|G^*_t=1, T=t) - E(Y+\delta^*_{Dt}|G^*_t=1, T=t-1)}{E(D|G^*_t=1, T=t) - E(D|G^*_t=1, T=t-1)} \\
&W^*_{TC}(-1,0,t) = \frac{E(Y|G^*_t=-1, T=t) - E(Y+\delta^*_{Dt}|G^*_t=-1, T=t-1)}{E(D|G^*_t=-1, T=t) - E(D|G^*_t=-1, T=t-1)}.
\end{aligned}$$

Finally, let

$$Q^*_{dt}(y) = F^{-1}_{Y|D=d,G^*_t=0,T=t} \circ F_{Y|D=d,G^*_t=0,T=t-1}(y) \quad d \in \{0,1\}$$
$$W^*_{CIC}(1,0,t) = \frac{E(Y|G^*_t=1,T=t) - E(Q^*_{Dt}(Y)|G^*_t=1,T=t-1)}{E(D|G^*_t=1,T=t) - E(D|G^*_t=1,T=t-1)}$$
$$W^*_{CIC}(-1,0,t) = \frac{E(Y|G^*_t=-1,T=t) - E(Q^*_{Dt}(Y)|G^*_t=-1,T=t)}{E(D|G^*_t=-1,T=t) - E(D|G^*_t=-1,T=t-1)}.$$

Theorem 4.1 *Assume that Model (1) and Assumption 1 are satisfied. Assume also that for every $t \in \{1,...,\bar{t}\}$, $\mathcal{G}_{st} \neq \emptyset$. Finally, assume that $G \perp\!\!\!\perp T$.*

1. *If Assumptions 3M and 4M are satisfied,*

$$\sum_{t=1}^{\bar{t}} w_t(w_{10|t} W^*_{DID}(1,0,t) + (1-w_{10|t}) W^*_{DID}(-1,0,t)) = \Delta_w.$$

2. *If Assumption 5M is satisfied,*

$$\sum_{t=1}^{\bar{t}} w_t(w_{10|t} W^*_{TC}(1,0,t) + (1-w_{10|t}) W^*_{TC}(-1,0,t)) = \Delta_w.$$

3. *If Assumptions 6 and 7 are satisfied,*

$$\sum_{t=1}^{\bar{t}} w_t(w_{10|t} W^*_{CIC}(1,0,t) + (1-w_{10|t}) W^*_{CIC}(-1,0,t)) = \Delta_w.$$

4. *If either $\bar{t} = 1$ or Assumption 11 holds,*

$$\Delta_w = E(Y(1) - Y(0)|S, T > 0).$$

Let us first consider the simple case with multiple groups and two periods. In such instances, the first, second, and third results of the theorem can respectively be rewritten as

$$\begin{aligned}
w_{10|1} W^*_{DID}(1,0,1) + (1-w_{10|1}) W^*_{DID}(-1,0,1) &= E(Y(1) - Y(0)|S_1, T = 1), \\
w_{10|1} W^*_{TC}(1,0,1) + (1-w_{10|1}) W^*_{TC}(-1,0,1) &= E(Y(1) - Y(0)|S_1, T = 1), \\
w_{10|1} W^*_{CIC}(1,0,1) + (1-w_{10|1}) W^*_{CIC}(-1,0,1) &= E(Y(1) - Y(0)|S_1, T = 1).
\end{aligned}$$

This shows that with multiple groups and two periods of time, treatment effects for switchers are identified if there is at least one group in which the treatment rate is stable over time. This holds under each of the three sets of assumptions we considered in the previous section. The estimands we propose can be computed in four steps. First, we form three "super groups", by pooling together the groups where treatment increases ($G^* = 1$), those where it is stable ($G^* = 0$), and those where it decreases ($G^* = -1$). While in some applications these three

sets of groups are known to the analyst, in other applications they must be estimated. In our supplementary material, we review results from Gentzkow et al. (2011) where these sets are known to the analyst. In Section 6 we review results from Duflo (2001) where these sets are not known to the analyst and need to be estimated. Second, we compute the estimand we suggested in the previous section with $G^* = 1$ and $G^* = 0$ as the treatment and control groups. Third, we compute the estimand we suggested in the previous section with $G^* = -1$ and $G^* = 0$ as the treatment and control groups. Finally, we compute a weighted average of those two estimands.

In the general case where $\bar{t} > 1$, aggregating estimands at different dates proves more difficult than aggregating estimands from different groups. This is because populations switching treatment between different dates might overlap. For instance, if a unit goes from non treatment to treatment between period 0 and 1, and from treatment to non treatment between period 1 and 2, she both belongs to period 1 and period 2 switchers. A weighted average of, say, our Wald-DID estimands between period 0 and 1 and between period 1 and 2 estimates a weighted average of the LATEs of two potentially overlapping populations. There is therefore no natural way to weight these two estimands to recover the LATE of the union of period 1 and 2 switchers. As shown in the fourth point of the theorem, the aggregated estimand we put forward still satisfies a nice property: it is equal to the LATE of the union of switchers in the special case where each group experiences a monotonic evolution of its treatment rate over time. When this is the case, populations switching treatment status at different dates cannot overlap, so our weighted average of switchers' LATE across periods is actually the LATE of all switchers.

Theorem 4.1 relies on the Assumption that $G \perp\!\!\!\perp T$. This requires that the distribution of groups be stable over time. This will automatically be satisfied if the data is a balanced panel and G is time invariant. With repeated cross-sections or cohort data, this assumption might fail to hold. However, large deviations from this stable group assumption indicate that some groups grow much faster than others, which might anyway call into question the common trends assumptions underlying DID identification strategies. Moreover, this assumption is only a sufficient condition to rationalize our estimands under assumptions at the group level. Another way to rationalize our estimands is to state our assumptions directly at the "super group" level. For instance, if Assumptions 1, 3M, and 5M are satisfied with G_t^* instead of G, then the first statement of Theorem 4.1 is still valid even if G is not independent of T. Finally, when G is not independent of T, it is still possible to form a Wald-DID and a Wald-TC type of estimand identifying a weighted average of LATEs under group-level assumptions. To do so, one merely needs to implement some reweighting to ensure that the distribution of groups is the same in periods $t-1$ and t in the reweighted population. For all $(g,t) \in \{0,1,...,\overline{g}\} \times \{1,...,\bar{t}\}$, let

$$r_{gt} = \frac{P(G=g|T=t)}{P(G=g|T=t-1)}.$$

One can show that a weighted average of

$$\frac{E(Y|G_t^*=1,T=t)-E\left(r_{Gt}Y|G_t^*=1,T=t-1\right)-\left(E(Y|G_t^*=0,T=t)-E\left(r_{Gt}Y|G_t^*=0,T=t-1\right)\right)}{E(D|G_t^*=1,T=t)-E\left(r_{Gt}D|G_t^*=1,T=t-1\right)-\left(E(D|G_t^*=0,T=t)-E\left(r_{Gt}D|G_t^*=0,T=t-1\right)\right)}$$

and

$$\frac{E(Y|G_t^*=-1,T=t)-E\left(r_{Gt}Y|G_t^*=-1,T=t-1\right)-\left(E(Y|G_t^*=0,T=t)-E\left(r_{Gt}Y|G_t^*=0,T=t-1\right)\right)}{E(D|G_t^*=-1,T=t)-E\left(r_{Gt}D|G_t^*=-1,T=t-1\right)-\left(E(D|G_t^*=0,T=t)-E\left(r_{Gt}D|G_t^*=0,T=t-1\right)\right)}$$

identifies a weighted average of LATEs under Assumptions 1, 3M, and 4M even if G is not independent of T.[10] One can follow similar steps to construct a Wald-TC type of estimand identifying a weighted average of LATEs under Assumptions 1 and 5M even if G is not independent of T.

Three last comments on Theorem 4.1 are in order. First, it contrasts with the current practice in empirical work. When many groups and periods are available, researchers usually include group and time fixed effects in their regressions, instead of pooling together groups into super control and treatment groups as we advocate here. In de Chaisemartin & D'Haultfœuille (2015), we show that such regressions estimate a weighted average of switchers' LATEs across groups, with potentially many negative weights and without the aggregation property we obtain here (see Theorems S1 and S2). Second, groups where the treatment rate diminishes can be used as "treatment" groups, just as those where it increases. Indeed, it is easy to show that all the results from the previous section still hold if the treatment rate decreases in the treatment group and is stable in the control group. Finally, when there are more than two groups where the treatment rate is stable between two consecutive dates, our three sets of assumptions become testable. Under each set of assumptions, using any subset of $\mathcal{G}_{st}$ as the control group should yield the same result.

We now turn to partial identification results when the treatment rate changes in every group. To simplify the exposition, we focus on the case with multiple groups and two periods. Results can easily be extended to accommodate multiple periods.

When the outcome has bounded support $[\underline{y},\overline{y}]$, let, for $(d,g)\in\{0,1\}\times\mathcal{S}(G)$,

$$\begin{aligned}
&\underline{F}_{dg1}(y)=M_0\left[1-\lambda_{gd}(1-F_{Y_{dg1}}(y))\right]-M_0(1-\lambda_{gd})\mathbb{1}\{y<\overline{y}\},\\
&\overline{F}_{dg1}(y)=m_1\left[\lambda_{gd}F_{Y_{dg1}}(y)\right]+(1-m_1(\lambda_{gd}))\mathbb{1}\{y\geq\underline{y}\}.
\end{aligned}$$

Then define

$$\begin{aligned}
&\overline{\delta}_d^-=\max_{g\in\mathcal{S}(G)}\int y d\overline{F}_{dg1}(y)-E(Y_{dg0}),\ \underline{\delta}_d^+=\min_{g\in\mathcal{S}(G)}\int y d\underline{F}_{dg1}(y)-E(Y_{dg0}),\\
&W_{TC}^-(g)=\frac{E(Y_{g1})-E(Y_{g0}+\underline{\delta}_{D_{g0}}^+)}{E(D_{g1})-E(D_{g0})},\ W_{TC}^+(g)=\frac{E(Y_{g1})-E(Y_{g0}+\overline{\delta}_{D_{g0}}^-)}{E(D_{g1})-E(D_{g0})}.
\end{aligned}$$

[10] The weights are the same as those in Theorem 4.1, except that one needs to replace $P(G_t^*=1)$ and $P(G_t^*=-1)$ by $P(G_t^*=1|T=t)$ and $P(G_t^*=-1|T=t)$ in their definition.

Let also $\underline{F}_{gg'd}(y)$ and $\overline{F}_{gg'd}(y)$ denote the lower and upper bounds on $F_{Y_{g1}(d)|S_g}$ one can obtain using $G = g$ as the treatment group and $G = g'$ as the control group and applying Theorem 3.3. Finally, let

$$W^-_{CIC}(g) = \int \left(\max_{g' \in \mathcal{S}(G)} \underline{F}_{gg'0}(y) - \min_{g' \in \mathcal{S}(G)} \overline{F}_{gg'1}(y) \right) dy, \; W^+_{CIC}(g) \quad = \int \left(\min_{g' \in \mathcal{S}(G)} \overline{F}_{gg'0}(y) - \max_{g' \in \mathcal{S}(G)} \underline{F}_{gg'1}(y) \right) dy.$$

Theorem 4.2 *Assume that Model* (1) *and Assumption 1 is satisfied. Assume also that* $\mathcal{G}_{s1} = \emptyset$.

1. *If Assumption 5 is satisfied and* $P(\underline{y} \leq Y(d) \leq \overline{y}) = 1$ *for* $d \in \{0, 1\}$,

$$W^-_{TC}(g) \leq E(Y_{g1}(1) - Y_{g1}(0)|S_{g1}) \leq W^+_{TC}(g).$$

2. *If Assumptions 6 and 7 are satisfied,*

$$W^-_{CIC}(g) \leq E(Y_{g1}(1) - Y_{g1}(0)|S_{g1}) \leq W^+_{CIC}(g).$$

This theorem shows that with multiple groups, one can construct intersection bounds for switchers' LATE when the treatment rate changes in every group over time. This holds under the two sets of assumptions for which we considered partial identification results in the previous section. Under Assumption 5, one can bound the LATE among switchers in a given group by using every other group as a potential control group and applying Theorem 3.2. One can then select the control group yielding the highest (resp. smallest) lower (resp. upper) bound. Under Assumption 6, one can bound the cdf of $Y(1)$ and $Y(0)$ among switchers in a given group by using every other group as a potential control group and applying Theorem 3.3. For each value of y, one can then select the control group yielding the highest (resp. lowest) lower (resp. upper) bound. One can finally bound switchers LATEs by using integration by parts for Lebesgue-Stieljes integrals. Note that any group can be used to construct bounds for the LATE of switchers in group g, even groups g' which experienced a larger change of their treatment rate. Here, we only present partial identification results for treatment effects among switchers of group g. One can also derive bounds for the entire population of switchers, by taking a weighted average of these bounds.

4.2 Covariates

We now return to our initial setup with two groups and two periods but consider a framework incorporating covariates. Let X be a vector of covariates. Assume that

$$\begin{aligned} Y(d) &= h_d(U_d, T, X), \quad d \in \mathcal{S}(D), \\ D &= 1\{V \geq v_{GTX}\}, \quad v_{G0X} = v_{00X} \text{ does not depend on } G. \end{aligned} \tag{10}$$

Then we replace Assumptions 1-7 by the following conditions.

Assumption 1*X* *(Conditional time invariance within groups)*

For $d \in \mathcal{S}(D)$, $(U_d, V) \perp\!\!\!\perp T \,|\, G, X$.

Assumption 2*X* *(Conditional first stage)*

Almost surely, $E(D_{11}|X) > E(D_{10}|X)$, and $E(D_{11}|X) - E(D_{10}|X) > E(D_{01}|X) - E(D_{00}|X)$.

Assumption 3*X* *(Conditional common trends)*

Almost surely, $E(h_0(U_0, 1, X) - h_0(U_0, 0, X)|G, X)$ does not depend on G.

Assumption 4*X* *(Conditional common effect of time on both potential outcomes)*

Almost surely,

$$E(h_1(U_1, 1, X) - h_1(U_1, 0, X)|G, V \geq v_{00X}, X) = E(h_0(U_0, 1, X) - h_0(U_0, 0, X)|G, V \geq v_{00X}, X).$$

Assumption 5*X* *(Conditional common trends within treatment status)*

Almost surely, $E(h_0(U_0, 1, X) - h_0(U_0, 0, X)|G, V < v_{00X}, X)$ and $E(h_1(U_1, 1, X) - h_1(U_1, 0, X)|G, V \geq v_{00X}, X)$ do not depend on G.

Assumption 6*X* *(Monotonicity)*

$U_d \in \mathbb{R}$ and $h_d(u, t, x)$ is strictly increasing in u for all $(d, t, x) \in \mathcal{S}(D) \times \mathcal{S}(T) \times \mathcal{S}(X)$.

Assumption 7*X* *(Data restrictions)*

1. *$\mathcal{S}(Y_{dgt}|X = x) = \mathcal{S}(Y) = [\underline{y}, \overline{y}]$ with $-\infty \leq \underline{y} < \overline{y} \leq +\infty$, for $(d, g, t, x) \in \mathcal{S}(D) \times \mathcal{S}(G) \times \mathcal{S}(T) \times \mathcal{S}(X)$.*
2. *$F_{Y_{dgt}|X=x}$ is strictly increasing on $\mathbb{R}$ and continuous on $\mathcal{S}(Y)$, for $(d, g, t, x) \in \mathcal{S}(D) \times \mathcal{S}(G) \times \mathcal{S}(T) \times \mathcal{S}(X)$.*
3. *$\mathcal{S}(X_{gt}) = \mathcal{S}(X)$ for $(g, t) \in \mathcal{S}(G) \times \mathcal{S}(T)$.*

For any random variable R, let $DID_R(X) = E(R_{11}|X) - E(R_{10}|X) - (E(R_{01}|X) - E(R_{00}|X))$. We also let $\delta_d(x) = E(Y_{d01}|X = x) - E(Y_{d00}|X = x)$, $Q_{d,x}(y) = F^{-1}_{Y_{d01}|X=x} \circ F_{Y_{d00}|X=x}(y)$, and

$$\begin{aligned} W_{DID}(X) &= \frac{DID_Y(X)}{DID_D(X)} \\ W_{TC}(X) &= \frac{E(Y_{11}|X) - E(Y_{10} + \delta_{D_{10}}(X)|X)}{E(D_{11}|X) - E(D_{10}|X)} \\ W_{CIC}(X) &= \frac{E(Y_{11}|X) - E(Q_{D_{10},X}(Y_{10})|X)}{E(D_{11}|X) - E(D_{10}|X)}. \end{aligned}$$

Finally, let $S_1 = \{V \in [v_{11X}, v_{00X}), G = 1\}$ and $\Delta(X) = E(Y_{11}(1) - Y_{11}(0)|S_1, X)$.

Theorem 4.3 *Assume that Model* (10) *and Assumptions 1X-2X hold, and that for every* $d \in \mathcal{S}(D)$, $0 < P(D_{00} = d|X) = P(D_{01} = d|X)$ *almost surely. Then*

1. *If Assumptions 3X-4X are satisfied,* $W_{DID}(X) = \Delta(X)$ *and*

$$W^X_{DID} \equiv \frac{E[DID_Y(X)|G=1,T=1]}{E[DID_D(X)|G=1,T=1]} = \Delta.$$

2. *If Assumption 5X is satisfied,* $W_{TC}(X) = \Delta(X)$ *and*

$$W^X_{TC} \equiv \frac{E(Y_{11}) - E[E(Y_{10} + D_{10}\delta_1(X) + (1-D_{10})\delta_0(X)|X)|G=1,T=1]}{E(D_{11}) - E(E(D_{10}|X)|G=1,T=1)} = \Delta.$$

3. *If Assumptions 6X-7X are satisfied,* $W_{CIC}(X) = \Delta(X)$ *and*

$$W^X_{CIC} \equiv \frac{E(Y_{11}) - E[E(D_{10}Q_{1,X}(Y_{10}) + (1-D_{10})Q_{0,X}(Y_{10})|X)|G=1,T=1]}{E(D_{11}) - E(E(D_{10}|X)|G=1,T=1)} = \Delta.$$

Incorporating covariates into the analysis has two advantages. First, it allows us to weaken our identifying assumptions. For instance, when the distribution of some X evolves over time in the control or in the treatment group, Assumption 1X is more plausible than Assumption 1: if the distribution of X is not stable over time and X is correlated with (U_d, V), then the distribution of (U_d, V) is also not stable. Second, there might be instances where $P(D_{00} = d) \neq P(D_{01} = d)$ but $P(D_{00} = d|X) = P(D_{01} = d|X) > 0$ almost surely, meaning that in the control group, the evolution of the treatment rate is entirely driven by a change in the distribution of X. If that is the case, one can use the previous theorem to point identify treatment effects among switchers, while our theorems without covariates only yield bounds. When $P(D_{00} = d|X) \neq P(D_{01} = d|X)$, one can derive bounds for $\Delta(X)$ and then for Δ. These bounds could be tighter than the unconditional ones if changes in the distribution of X drive most of the evolution of the treatment rate in the control group.

4.3 Non-binary, ordered treatment

We first consider the case where the treatment is not binary but takes a finite number of values and is ordered: $D \in \{0, 1, ..., \overline{d}\}$. One prominent example is years of schooling, as in our application in Section 6. We extend our model to this case as follows:

$$\begin{array}{rcll} Y(d) & = & h_d(U_d, T), & \text{for } d \in \{0, ..., \overline{d}\}, \\ D & = & \sum_{d=1}^{\overline{d}} 1\{V \geq v^d_{GT}\}, & -\infty = v^0_{gt} < v^1_{gt} ... < v^{\overline{d}+1}_{gt} = +\infty \text{ for } (g,t) \in \{0,1\}^2. \end{array} \tag{11}$$

Assumption 4*O* *(Common average effect of time on all potential outcomes)*

For $d \in \{0, ..., \overline{d}\}$,

$$E(h_d(U_d, 1) - h_d(U_d, 0)|G, V \in [v^d_{G0}, v^{d+1}_{G0})) = E(h_0(U_0, 1) - h_0(U_0, 0)|G, V \in [v^d_{G0}, v^{d+1}_{G0})).$$

Assumption 5*O* *(Common trends within treatment status at date 0)*

For every $d \in \mathcal{S}(D)$, $E(h_d(U_d,1) - h_d(U_d,0)|G, V \in [v^d_{G0}, v^{d+1}_{G0}))$ *does not depend on* G.

Model (11) and Assumptions 4O-5O generalize respectively Model (1) and Assumptions 4-5 to situations where the treatment is non-binary and ordered . Let $\gtrsim$ denote stochastic dominance between two random variables, while $\sim$ denotes equality in distribution.

Theorem 4.4 *Assume that Model* (11) *and Assumptions 1-2 are satisfied, that* $D_{01} \sim D_{00}$, *and that* $D_{11} \gtrsim D_{10}$. *Let* $w_d = \frac{P(D_{11} \geq d) - P(D_{10} \geq d)}{E(D_{11}) - E(D_{10})}$.

1. *If Assumptions 3 and 4O are satisfied,*

$$W_{DID} = \sum_{d=1}^{\overline{d}} E(Y_{11}(d) - Y_{11}(d-1)|V \in [v^d_{11}, v^d_{10}))w_d.$$

2. *If Assumption 5O is satisfied,*

$$W_{TC} = \sum_{d=1}^{\overline{d}} E(Y_{11}(d) - Y_{11}(d-1)|V \in [v^d_{11}, v^d_{10}))w_d.$$

3. *If Assumptions 6 and 7 are satisfied,*

$$W_{CIC} = \sum_{d=1}^{\overline{d}} E(Y_{11}(d) - Y_{11}(d-1)|V \in [v^d_{11}, v^d_{10}))w_d.$$

Theorem 4.4 shows that with an ordered treatment, the estimands we considered in the previous sections are equal to the average causal response (ACR) parameter considered in Angrist & Imbens (1995). This parameter is a weighted average, over all values of d, of the effect of increasing treatment from $d-1$ to d among switchers whose treatment status goes from strictly below to above d over time.

For this theorem to hold, two conditions have to be satisfied. First, in the treatment group, the distribution of treatment in period 1 should dominate stochastically the corresponding distribution in period 0. Angrist & Imbens (1995) also require that the distribution of treatment conditional on $Z = 1$ dominate that conditional on $Z = 0$. Actually, this assumption is not necessary for our three estimands to identify a weighted sum of treatment effects. If it is not satisfied, one still has that W_{DID}, W_{TC}, or W_{CIC} identify

$$\sum_{d=1}^{\overline{d}} E(Y_{11}(d) - Y_{11}(d-1)|V \in [\min(v^d_{10}, v^d_{11}), \max(v^d_{10}, v^d_{11}))w_d,$$

which is a weighted sum of treatment effects with some negative weights. Second, the distribution of treatment should be stable over time in the control group. When it is not, one

can still obtain some identification results. Firstly, Theorem 3.1 generalizes to non-binary and ordered treatments taking a finite number of values. When treatment increases in the control group, the Wald-DID identifies a weighted difference of the ACRs in the treatment and in the control group; when treatment decreases in the control group, the Wald-DID identifies a weighted average of these two ACRs. The weights are the same as those in Theorem 3.1. Secondly, the second statement of Theorems 3.2 and 3.3 also generalize to non-binary and ordered treatments taking a finite number of values. When the distribution of treatment is not stable over time in the control group, the ACR in the treatment group can be bounded under Assumption 5O, or Assumptions 6 and 7.

Theorem 4.4 could easily be extended to continuous treatments. Our three estimators would then estimate a weighted average derivative similar to that studied in Angrist et al. (2000). However, non-parametric estimation of the Wald-CIC might be challenging, as one would have to estimate the function $d \mapsto Q_d$ in a first step.

5 Inference

In this section, we study the asymptotic properties of the estimators corresponding to the estimands introduced in the previous sections. We focus on the point identified case. Estimators of the bounds on average and quantile treatment effects in the partially identified case are considered in de Chaisemartin & D'Haultfœuille (2015). We restrict ourselves to repeated cross sections. For now, we suppose that an i.i.d. sample with the same distribution as (Y, D, G, T, X) is available.

Assumption 12 *(Independent and identically distributed observations)*

$(Y_i, D_i, G_i, T_i, X_i)_{i=1,...,n}$ *are i.i.d.*

Even if we do not observe the same unit twice, independence may be a strong assumption in some applications: clustering at the group level can induce both cross-sectional and serial correlation within clusters. However, we can extend some of our results to allow for clustering, as we discuss below.

5.1 Inference without covariates

Let $\mathcal{I}_{gt} = \{i : G_i = g, T_i = t\}$ (resp. $\mathcal{I}_{dgt} = \{i : D_i = d, G_i = g, T_i = t\}$) and n_{gt} (resp. n_{dgt}) denote the size of $\mathcal{I}_{gt}$ (resp. $\mathcal{I}_{dgt}$) for all $(d, g, t) \in \{0, 1\}^3$. The Wald-DID and Wald-TC

estimators are simply defined by

$$\widehat{W}_{DID} = \frac{\frac{1}{n_{11}}\sum_{i\in\mathcal{I}_{11}} Y_i - \frac{1}{n_{10}}\sum_{i\in\mathcal{I}_{10}} Y_i - \frac{1}{n_{01}}\sum_{i\in\mathcal{I}_{01}} Y_i + \frac{1}{n_{00}}\sum_{i\in\mathcal{I}_{00}} Y_i}{\frac{1}{n_{11}}\sum_{i\in\mathcal{I}_{11}} D_i - \frac{1}{n_{10}}\sum_{i\in\mathcal{I}_{10}} D_i - \frac{1}{n_{01}}\sum_{i\in\mathcal{I}_{01}} D_i + \frac{1}{n_{00}}\sum_{i\in\mathcal{I}_{00}} D_i},$$
$$\widehat{W}_{TC} = \frac{\frac{1}{n_{11}}\sum_{i\in\mathcal{I}_{11}} Y_i - \frac{1}{n_{10}}\sum_{i\in\mathcal{I}_{10}} \left[Y_i + \widehat{\delta}_{D_i}\right]}{\frac{1}{n_{11}}\sum_{i\in\mathcal{I}_{11}} D_i - \frac{1}{n_{10}}\sum_{i\in\mathcal{I}_{10}} D_i},$$

where $\widehat{\delta}_d$ ($d \in \{0,1\}$) is defined by

$$\widehat{\delta}_d = \frac{1}{n_{d01}}\sum_{i\in\mathcal{I}_{d01}} Y_i - \frac{1}{n_{d00}}\sum_{i\in\mathcal{I}_{d00}} Y_i.$$

Let $\widehat{F}_{Y_{dgt}}$ denote the empirical cdf of Y on the subsample $\mathcal{I}_{dgt}$:

$$\widehat{F}_{Y_{dgt}}(y) = \frac{1}{n_{dgt}}\sum_{i\in\mathcal{I}_{dgt}} \mathbb{1}\{Y_i \leq y\}.$$

Similarly, we estimate the quantile of order $q \in (0,1)$ of Y_{dgt} by $\widehat{F}^{-1}_{Y_{dgt}}(q) = \inf\{y : \widehat{F}_{Y_{dgt}}(y) \geq q\}$. The estimator of the quantile-quantile transform is $\widehat{Q}_d = \widehat{F}^{-1}_{Y_{d01}} \circ \widehat{F}_{Y_{d00}}$. Then, the Wald-CIC estimator is defined by

$$\widehat{W}_{CIC} = \frac{\frac{1}{n_{11}}\sum_{i\in\mathcal{I}_{11}} Y_i - \frac{1}{n_{10}}\sum_{i\in\mathcal{I}_{10}} \widehat{Q}_{D_i}(Y_i)}{\frac{1}{n_{11}}\sum_{i\in\mathcal{I}_{11}} D_i - \frac{1}{n_{10}}\sum_{i\in\mathcal{I}_{10}} D_i}.$$

Let $\widehat{P}(D_{gt} = d)$ be the proportion of subjects with $D = d$ in the sample $\mathcal{I}_{gt}$, let $\widehat{H}_d = \widehat{F}_{Y_{d10}} \circ \widehat{F}^{-1}_{Y_{d00}}$, and let

$$\widehat{F}_{Y_{11}(d)|S_1} = \frac{\widehat{P}(D_{10} = d)\widehat{H}_d \circ \widehat{F}_{Y_{d01}} - \widehat{P}(D_{11} = d)\widehat{F}_{Y_{d11}}}{\widehat{P}(D_{10} = d) - \widehat{P}(D_{11} = d)}.$$

Our estimator of the LQTE of order q for switchers is

$$\widehat{\tau}_q = \widehat{F}^{-1}_{Y_{11}(1)|S_1}(q) - \widehat{F}^{-1}_{Y_{11}(0)|S_1}(q).$$

We derive the asymptotic behavior of our CIC estimators under the following assumption, which is similar to the one made by Athey & Imbens (2006) for the CIC estimators in sharp settings.

Assumption 13 *(Regularity conditions for the CIC estimators)*

$\mathcal{S}(Y)$ is a bounded interval $[\underline{y}, \overline{y}]$. Moreover, for all $(d,g,t) \in \{0,1\}^3$, $F_{Y_{dgt}}$ and $F_{Y_{11}(d)|S_1}$ are continuously differentiable with strictly positive derivatives on $[\underline{y}, \overline{y}]$.

Theorem 5.1 below shows that all our estimators are root-n consistent and asymptotically normal. We also derive the influence functions of our estimators. However, because these influence functions take complicated expressions, using the bootstrap might be convenient for inference. For any statistic T, we let T^* denote its bootstrap counterpart. For any root-n consistent statistic $\widehat{\theta}$ estimating consistently θ, we say that the bootstrap is consistent if with probability one and conditional on the sample, $\sqrt{n}(\widehat{\theta}^*-\widehat{\theta})$ converges to the same distribution as the limit distribution of $\sqrt{n}(\widehat{\theta}-\theta)$.[11] Theorem 5.1 implies that bootstrap confidence intervals are asymptotically valid for all our estimators.

Theorem 5.1 *Suppose that Assumptions 1-2, 12 hold and* $0 < P(D_{00} = 1) = P(D_{01} = 1) < 1$. *Then*

1. *If* $E(Y^2) < \infty$ *and Assumptions 3-4 also hold,*

$$\sqrt{n}\left(\widehat{W}_{DID} - \Delta\right) \xrightarrow{L} \mathcal{N}\left(0, V\left(\psi_{DID}\right)\right),$$

   where ψ_{DID} *is defined in Equation* (42) *in the appendix. Moreover, the bootstrap is consistent for* $\widehat{W}_{DID}$.

2. *If* $E(Y^2) < \infty$ *and Assumption 5 also holds,*

$$\sqrt{n}\left(\widehat{W}_{TC} - \Delta\right) \xrightarrow{L} \mathcal{N}\left(0, V\left(\psi_{TC}\right)\right)$$

   where ψ_{TC} *is defined in Equation* (43) *in the appendix. Moreover, the bootstrap is consistent for* $\widehat{W}_{TC}$.

3. *If Assumptions 6, 7 and 13 also hold,*

$$\sqrt{n}\left(\widehat{W}_{CIC} - \Delta\right) \xrightarrow{L} \mathcal{N}\left(0, V\left(\psi_{CIC}\right)\right),$$
$$\sqrt{n}\left(\widehat{\tau}_q - \tau_q\right) \xrightarrow{L} \mathcal{N}\left(0, V\left(\psi_{q,CIC}\right)\right),$$

   where ψ_{CIC} *and* $\psi_{q,CIC}$ *are defined in Equations* (44) *and* (45) *in the appendix. Moreover, the bootstrap is consistent for both estimators.*

The result is straightforward for the Wald-DID and Wald-TC. Regarding the CIC, our proof differs from the one of Athey & Imbens (2006). It is based on the weak convergence of the empirical cdfs of the different subgroups, and on a repeated use of the functional delta method. This approach can be readily applied to other functionals of $(F_{Y_{11}(0)|S_1}, F_{Y_{11}(1)|S_1})$. We also show in the supplementary material how it can be applied to estimate bounds on average and quantile treatment effects in the partially identified case.

[11] See, e.g., van der Vaart (2000), Section 23.2.1, for a formal definition of conditional convergence.

5.2 Inference with covariates

In this section, we consider estimators of the Wald-DID, Wald-TC, and Wald-CIC estimands with covariates derived in Subsection 4.2. For the Wald-DID and Wald-TC, our estimators are entirely non-parametric.[12] For the Wald-CIC, we could define an estimator using a non-parametric estimator of the conditional quantile-quantile transform $Q_{d,X}$. However, such an estimator would be cumbersome to compute. Following Melly & Santangelo (2015), we consider instead an estimator of $Q_{d,X}$ based on quantile regressions. This estimator relies on the assumption that conditional quantiles of the outcome are linear. However, it does not require that the effect of the treatment be the same for units with different values of their covariates, contrary to the estimator with covariates suggested in Athey & Imbens (2006).

Let us assume that $X \in \mathbb{R}^r$ is a vector of continuous covariates. Adding discrete covariates is easy by reasoning conditional on each corresponding cell. We take an approach similar to, e.g., Frölich (2007) by estimating in a first step conditional expectations by series estimators. For any positive integer K, let $p^K(x) = (p_{1K}(x), ..., p_{KK}(x))'$ be a vector of basis functions and $P^K = (p^K(X_1), ..., p^K(X_n))$. For any random variable R, we estimate $m^R(x) = E(R|X = x)$ by the series estimator

$$\widehat{m}^R(x) = p^{K_n}(x)' \left(P^{K_n} P^{K_n\prime}\right)^- P^{K_n} (R_1, ..., R_n)',$$

where $(.)^-$ denotes the generalized inverse and $(K_n)_{n\in\mathbb{N}}$ is a sequence of integers tending to infinity at a rate specified below. Following Frölich (2007), for any $(g,t) \in \{0,1\}^2$ we estimate $m^R_{gt}(x) = E(R_{gt}|X = x)$ by $\widehat{m}^R_{gt}(x) = \widehat{m}^{\mathbb{1}\{G=g,T=t\}R}(x)/\widehat{m}^{\mathbb{1}\{G=g,T=t\}}(x)$. $m^R_{dgt}(x) = E(R_{dgt}|X = x)$ is estimated similarly. Then our Wald-DID and Wald-TC estimators with covariates are defined by

$$\widehat{W}^X_{DID} = \frac{\frac{1}{n_{11}} \sum_{i\in\mathcal{I}_{11}} \left[Y_i - \widehat{m}^Y_{10}(X_i) - \widehat{m}^Y_{01}(X_i) + \widehat{m}^Y_{00}(X_i)\right]}{\frac{1}{n_{11}} \sum_{i\in\mathcal{I}_{11}} \left[D_i - \widehat{m}^D_{10}(X_i) - \widehat{m}^D_{01}(X_i) + \widehat{m}^D_{00}(X_i)\right]},$$

$$\widehat{W}^X_{TC} = \frac{\frac{1}{n_{11}} \sum_{i\in\mathcal{I}_{11}} \left[Y_i - \widehat{m}^Y_{10}(X_i) - \widehat{m}^D_{10}(X_i)\widehat{\delta}_1(X_i) - (1 - \widehat{m}^D_{10}(X_i))\widehat{\delta}_0(X_i)\right]}{\frac{1}{n_{11}} \sum_{i\in\mathcal{I}_{11}} \left[D_i - \widehat{m}^D_{10}(X_i)\right]},$$

where $\widehat{\delta}_d(x) = \widehat{m}^Y_{d01}(x) - \widehat{m}^Y_{d00}(x)$.

We then introduce our Wald-CIC estimator with covariates. Suppose that for all $(d,g,t,\tau) \in \{0,1\}^3 \times (0,1)$,

$$F_{Y_{dgt}|X=x} = x'\beta_{dgt}(\tau).$$

[12] In our Stata package, we also implement estimators relying on the assumption that all the conditional expectations in W^X_{DID} and W^X_{TC} are linear functions of X and can therefore be estimated through simple OLS regressions. These estimators might prove useful when the set of covariates is rich and the estimation of our non-parametric estimators is cumbersome. Asymptotic normality of these estimators follows directly from standard results on OLS regressions and the Delta method.

Using the fact that $F_{Y_{dgt}|X=x} = \int_0^1 \mathbb{1}\{F^{-1}_{Y_{dgt}|X=x}(\tau) \leq y\}d\tau$ (see, e.g., Chernozhukov et al., 2010), we obtain

$$Q_{d,x}(y) = x'\beta_{d01}\left(\int_0^1 \mathbb{1}\{x'\beta_{d00}(\tau) \leq y\}d\tau\right).$$

Besides, some algebra shows that

$$E\left[Q_{D_{10},X}(Y_{10})|X\right] = m^D_{10}(X)\int_0^1 Q_{1,X}(X'\beta_{110}(u))du + (1-m^D_{10}(X))\int_0^1 Q_{0,X}(X'\beta_{010}(u))du.$$

Hence, we estimate $\widehat{W}^X_{CIC}$ by

$$\widehat{W}^X_{CIC} = \frac{\frac{1}{n_{11}}\sum_{i\in\mathcal{I}_{11}}\left[Y_i - \widehat{m}^D_{10}(X_i)\int_0^1 \widehat{Q}_{1,X_i}(X_i'\widehat{\beta}_{110}(u))du - (1-\widehat{m}^D_{10}(X_i))\int_0^1 \widehat{Q}_{0,X_i}(X_i'\widehat{\beta}_{010}(u))du\right]}{\frac{1}{n_{11}}\sum_{i\in\mathcal{I}_{11}}\left[D_i - \widehat{m}^D_{10}(X_i)\right]},$$

where the estimator of the conditional quantile-quantile transform satisfies

$$\widehat{Q}_{d,x}(y) = x'\widehat{\beta}_{d01}\left(\int_0^1 \mathbb{1}\{x'\widehat{\beta}_{d00}(\tau) \leq y\}d\tau\right),$$

and $\widehat{\beta}_{dgt}(\tau)$ is obtained from a quantile regression of Y on X on the subsample $\mathcal{I}_{dgt}$:

$$\widehat{\beta}_{dgt}(\tau) = \arg\min_{\beta\in B}\sum_{i\in\mathcal{I}_{dgt}}(\tau - \mathbb{1}\{Y_i - X_i'\beta \leq 0\})(Y_i - X_i'\beta).$$

Here B denotes a compact subset of $\mathbb{R}^r$ including $\beta_{dgt}(\tau)$ for all $(d,g,t,\tau) \in \{0,1\}^3 \times (0,1)$. In practice, instead of computing the whole quantile regression process, we can compute $\tau \mapsto \widehat{\beta}_{dgt}(\tau)$ on a fine enough grid and replace integrals by corresponding averages. See Melly & Santangelo (2015) for a detailed discussion on computational issues.

We prove the asymptotic normality of our estimators under the following assumptions.

Assumption 14 *(Regularity conditions for the series estimators)*

1. *For any $(d,g,t,\alpha) \in \{0,1\}^3\times\{0,1,2\}$, $\inf_{x\in\mathcal{S}(X)} P(D=d,G=g,T=t|X=x) > 0$ and $x \mapsto E(\mathbb{1}\{D=d\}\mathbb{1}\{G=g\}\mathbb{1}\{T=t\}Y^\alpha|X=x)$ is s times continuously differentiable on $\mathcal{S}(X)$, with $s > 3r$.*

2. *$\mathcal{S}(X)$ is a Cartesian product of compact connected intervals on which X has a probability density function that is bounded away from zero. Moreover $E(XX')$ is nonsingular.*

3. *The series terms p_{kK_n}, $1 \leq k \leq K_n$, are products of polynomials orthonormal with respect to the uniform weight. Moreover, $K_n^{4(s/r-1)}/n \to \infty$ and $K_n^7/n \to 0$.*

Assumption 15 *(Regularity conditions for the conditional Wald-CIC estimator)*

For all $(d,g,t,x,\tau) \in \{0,1\}^3 \times \mathcal{S}(X) \times (0,1)$, $F^{-1}_{Y_{dgt}|X=x}(\tau) = x'\beta_{dgt}(\tau)$, with $\beta_{dgt}(\tau) \in B$, a compact subset of $\mathbb{R}^r$. Moreover, $F_{Y_{dgt}|X=x}$ is differentiable, with

$$0 < \inf_{(x,y)\in\mathcal{S}(X)\times\mathcal{S}(Y)} f_{Y_{dgt}|X=x}(y) \leq \sup_{(x,y)\in\mathcal{S}(X)\times\mathcal{S}(Y)} f_{Y_{dgt}|X=x}(y) < +\infty.$$

Assumption 15 implies that Y has a compact support. If its conditional density is not bounded away from zero, trimming may be necessary as discussed in Chernozhukov, Fernández-Val & Melly (2013) and Melly & Santangelo (2015).

Theorem 5.2 *Suppose that Model* (10) *and Assumptions 1X-2X, 12 and 14 hold. Then*

1. *If Assumptions 3X-4X also hold,*

$$\sqrt{n}\left(\widehat{W}^X_{DID} - \Delta\right) \xrightarrow{L} \mathcal{N}\left(0, V(\psi^X_{DID})\right),$$

where the variable ψ^X_{DID} is defined in Equation (46) *in the appendix.*

2. *If Assumption 5X also holds,*

$$\sqrt{n}\left(\widehat{W}^X_{TC} - \Delta\right) \xrightarrow{L} \mathcal{N}\left(0, V(\psi^X_{TC})\right),$$

where the variable ψ^X_{TC} is defined in Equation (47) *in the appendix.*

3. *If Assumptions 6X-7X and 15 also hold,*

$$\sqrt{n}\left(\widehat{W}^X_{CIC} - \Delta\right) \xrightarrow{L} \mathcal{N}\left(0, V(\psi^X_{CIC})\right),$$

where the variable ψ^X_{CIC} is defined in Equation (49) *in the appendix.*

We prove the asymptotic normality of the Wald-DID and Wald-TC estimators using repeatedly results on two-step estimators involving nonparametric first-step estimators, see e.g. Newey (1994). Proving the asymptotic normality of the Wald-CIC estimator is more challenging. We have to prove the weak convergence of $\sqrt{n}\left(\widehat{\beta}_{dgt}(.) - \beta_{dgt}(.)\right)$, seen as a stochastic process, on the whole interval $(0,1)$. To our knowledge, this convergence has been established so far only on $[\varepsilon, 1-\varepsilon]$, for any $\varepsilon > 0$ (see, e.g., Angrist et al., 2006). Here, this result holds thanks to our assumptions on the conditional distribution of Y. Finally, note that our Wald-CIC estimator does not require any first-step nonparametric estimator in the special case where $P(D_{10} = 1) = 0$. In such a case, asymptotic normality still holds without the regularity conditions in Assumption 14. Only the nonsingularity of $E(XX')$ is needed. In our supplementary material, we revisit results from Field (2007), where $P(D_{10} = 1) = 0$ and where the set of covariates is very rich.

5.3 Accounting for clustering

In many applications, the i.i.d. condition in Assumption 12 is too strong, because of cross-sectional or serial dependence within clusters. However, in such instances one can build upon our previous results to draw inference on the Wald-DID and Wald-TC without covariates, and on the Wald-CIC without covariates if clusters are of the same size.

We consider an asymptotic framework where the number of clusters C tends to infinity while the sample size within each cluster remains bounded in probability. Let $n_c = \#\{i \in c\}$, $\overline{n}_c = \frac{1}{C}\sum_{c=1}^{C} n_c$, $n_{ct} = \#\{i \in c : T_i = t\}$, $n_{cdt} = \#\{i \in c : T_i = t, D_i = d\}$, $D_{ct} = \frac{1}{n_{ct}}\sum_{i \in c:T_i=t} D_i$, $Y_{ct} = \frac{1}{n_{ct}}\sum_{i \in c:T_i=t} Y_i$, and $Y_{cdt} = \frac{1}{n_{cdt}}\sum_{i \in c:T_i=t,D_i=d} Y_i$, with the convention that the sums are equal to zero if they sum over empty sets. Then we can write the estimators of the Wald-DID and Wald-TC as simple functions of averages of these variables defined at the cluster level. Using the same reasoning as in the proof of Theorem 5.1, we can linearize both estimators, ending up with

$$\sqrt{C}\left(\widehat{W}_{DID} - \Delta\right) = \frac{1}{\sqrt{C}}\sum_{c=1}^{C} \frac{n_c}{\overline{n}_c}\psi_{c,DID} + o_P(1),$$

$$\sqrt{C}\left(\widehat{W}_{TC} - \Delta\right) = \frac{1}{\sqrt{C}}\sum_{c=1}^{C} \frac{n_c}{\overline{n}_c}\psi_{c,TC} + o_P(1),$$

where $\psi_{c,DID} = \frac{1}{n_c}\sum_{i \in c}\psi_{i,DID}$ and similarly for $\psi_{c,TC}$. In other words, to estimate the asymptotic variance of our estimators while accounting for clustering, it suffices to compute the average over clusters of the influence functions we obtained assuming that observations were i.i.d, multiply them by $\frac{n_c}{\overline{n}_c}$, and then compute the variance of this variable over clusters.

Our other estimators cannot be written as functions of variables aggregated at the cluster level: they depend on the variables of every unit in each cluster. But as long as they can still be linearized in the presence of clustering, the same argument as above applies. Such a linearization can be obtained for the Wald-CIC estimator with clusters of same size, because weak convergence of the empirical cdfs of the different subgroups still holds in this context.[13] We conjecture that it can also be obtained when clusters are of random sizes, or with our estimators including covariates. Proving this last point would nevertheless require to adapt results on two-step estimators to such a clustering framework. To the best of our knowledge, no such results have been established yet.

6 Application: returns to education in Indonesia

6.1 Estimation strategy

In 1973-1974, the Indonesian government launched a major primary school construction program, the so-called INPRES program. Duflo (2001) uses it to measure returns to education among men through a fuzzy DID identification strategy. In her analysis, groups are districts,

[13] To simplify, let us ignore the different subgroups and let us consider the standard empirical process on Y. Let $\mathbf{Y}_c = (Y_{c1}, ..., Y_{cn_c})'$, where Y_{ci} denotes the outcome variable of individual i in cluster c. Because the $(\mathbf{Y}_c)_{c=1...C}$ are i.i.d., its multivariate empirical process converges to a multivariate gaussian process. The standard empirical process on Y can be written as the average over the n_c components of this multivariate process. Therefore, it also converges to a gaussian process.

the administrative unit at which the program was implemented. This definition of groups could violate Assumption 1 if the program generated endogenous migration between districts. The author therefore uses district of birth instead of district of residence. She then constructs two "super groups" of treatment and control districts, by regressing the number of schools constructed on the number of school-age children in each district. Treatment districts are those with a positive residual in that regression, as they received more schools than what their population predicts. She also uses the fact that exposure to treatment varied across cohorts. Children born between 1968 and 1972 entered primary school after the program was launched, while children born between 1957 and 1962 had finished primary school by that time.

However, the INPRES program explains a small fraction of the differences in increases in years of schooling between districts. A district-level regression of the increase in years of schooling between these two groups of cohorts on the number of primary schools constructed per school-age children has an R-squared of 0.03 only. The INPRES program was not the only school construction program taking place at that time: between 1973 and 1983, the number of primary, middle, and high schools in the country respectively increased by 96, 94, and 139%. Including the change in the number of middle and high schools in the district-level regression increases its R-squared to 0.14, but still leaves most of the variation unexplained.

Because of this, the results in Duflo's paper rely on the assumption that returns to education are homogeneous between districts. The author first uses a simple Wald-DID with her two groups of districts and cohorts to estimate returns to education. Under Assumptions 1-3 and 4O, one can show that this simple Wald-DID is equal to $\frac{0.47}{0.11}ACR_1 - \frac{0.36}{0.11}ACR_0$, where ACR_1 and ACR_0 respectively denote the ACR parameters we introduced in Section 4.3 in the treatment and in the control group, and where the weights can be computed from Table 3.[14] If $ACR_1 \neq ACR_0$, this simple Wald-DID could lie far from both ACR_1 and ACR_0. Then, the author considers richer specifications. All of them include cohort and district of birth fixed effects. We show in the supplementary material (see Theorem S2) that such regressions estimate a weighted sum of switchers returns to education across districts, with potentially many negative weights. We estimate the weights received by each district in her data, and find that almost half of districts receive a negative weight, with negative weights summing up to -3.28. Here again, if switchers' returns are heterogeneous across districts with positive and negative weights, these regression coefficients could lie very far from returns in any district. Therefore, these richer specifications also rely on the assumption that returns to education are homogeneous across districts.

This assumption is not warranted in this context. As one can see in Table 3, educational attainment in the older cohort is substantially higher in control than in treatment districts, implying that the supply of skilled labor is higher there. Returns to education could be

[14] Theorem 3.1 can easily be generalized to non-binary, ordered treatments.

lower in control districts if the two groups face the same demand for skilled labor. On the other hand, this difference in educational attainment might also indicate a higher level of economic development in control districts, in which case demand for skilled labor and returns to education could be higher there.

Table 3: Average number of years of education completed

	Cohort 0	Cohort 1	Evolution	s.e.
Groups in Duflo (2001)				
Treatment districts	8.02	8.49	0.47	(0.070)
Control districts	9.40	9.76	0.36	(0.038)
New groups				
Treatment districts	8.65	9.64	0.99	(0.082)
Control districts	9.60	9.55	-0.05	(0.097)

Notes. This table reports the evolution of average years of schooling between cohorts 0 and 1 in the treatment and controls groups used by Duflo (2001) and in our new treatment and control groups. Standard errors are clustered at the district level.

To avoid relying on the assumption that treatment effects are homogeneous between districts, we use a different statistical procedure from that used by Duflo to classify districts into a treatment and a control group. This procedure should classify as controls only districts with a stable distribution of education. Any classification method leads us to make two types of errors: classify some districts where the distribution of education remained constant as treatments (type 1 error); and classify some districts where this distribution changed as controls (type 2 error). Type 1 errors are innocuous. For instance, if Assumptions 3 and 4O are satisfied, all control districts have the same evolution of their expected outcome. Misclassifying some as treatment districts leaves the Wald-DID estimator unchanged, up to sampling error. On the other hand, type 2 errors are a more serious concern. They lead us to include districts where the true distribution of education was not stable in our super control group, thus violating one of the requirements of Theorem 4.1.

We therefore choose a method based on chi-squared tests with very liberal level. Specifically, we assign a district to our control group if the p-value of a chi-squared test comparing the distribution of education between the two cohorts in that district is greater than 0.5. If that p-value is lower than 0.5 and the average number of years of education increased in that district, we assign it to our treatment group. We end up with control and treatment groups respectively made up of 64 and 123 districts. We exclude from the analysis 97 districts with a p-value lower than 0.5 and where years of education decreased. As shown in Section 4.1,

we could gather them together to form a third super group, and use results from Theorem 4.1 to estimate returns to education. However, doing this hardly changes our point estimates. We therefore stick to two super groups, to keep the presentation as simple as possible and to follow Duflo (2001) who also has two super treatment and control groups.

As shown in Table 3, in treatment districts the younger cohort completed one more year of education than the older one, while in control districts the two cohorts completed almost the same number of years of education. In treatment districts, the distribution of education in the younger cohort almost stochastically dominates that in the older cohort, as one can see from Table 4. The college completion rate is 2.5 percentage points higher in the older than in the younger cohort, but that difference is fairly small. Moreover, in control districts, the distribution of education is almost the same between the two cohorts. The primary school and college completion rate are respectively 2.6 percentage points higher and 3.3 percentage points lower in the younger cohort, but these differences are small too. Overall, the two requirements of Theorem 4.4 are close to being satisfied. We argue below that the minor departures from these two requirements that can be seen in Table 4 are unlikely to drive our results.

Table 4: Evolution of the distribution of education

	Cohort 0	Cohort 1	Evolution	s.e.
Treatment group				
Completed primary school	0.815	0.931	0.116	(0.008)
Completed middle school	0.531	0.676	0.145	(0.011)
Completed high school	0.406	0.491	0.085	(0.013)
Completed undergrad	0.094	0.069	-0.025	(0.006)
N	17471			
Control group				
Completed primary school	0.877	0.904	0.026	(0.008)
Completed middle school	0.640	0.656	0.016	(0.012)
Completed high school	0.510	0.489	-0.021	(0.013)
Completed undergrad	0.104	0.071	-0.033	(0.006)
N	4868			

Notes. This table reports the evolution of schooling between cohorts 0 and 1 by broad categories in our new treatment and control groups. Standard errors are clustered at the district level.

Finally, we consider two placebo experiments to assess the plausibility of the common trends assumptions underlying our estimators with our "super groups". First, following Duflo (2001),

we compare years of schooling and wages for men born between 1957 and 1962 and those born between 1951 and 1956 (cohort -1). Then, we compare men born between 1951 and 1956 and those born between 1945 and 1950 (cohort -2). Results lend strong support to our identification strategy. The difference in average years of education between the two groups of districts is stable in the three older cohorts, but it is much larger for the younger cohort. Accordingly, the difference in average wages between the two groups of districts is also very stable in the three older cohorts, but it is much larger for the younger cohort. This remains true when instead of comparing average wages we estimate the numerator of the Wald-TC and of the Wald-CIC. While the placebo estimators are small and insignificant, the true estimators are large and significant. Theorem 4.1 relies on the assumption that $G \perp\!\!\!\perp T$. This assumption fails to hold here: the distribution of districts is not perfectly stable between the two cohorts. However, our placebo tests suggest that our common trend assumptions are satisfied directly at the "super group" level, thus implying that deviations from $G \perp\!\!\!\perp T$ are not a serious concern for our results.

Table 5: Placebo tests

	Cohort -2 versus -1	Cohort -1 versus 0	Cohort 0 versus 1
DID schooling	0.108	-0.006	1.030
	(0.191)	(0.160)	(0.127)
DID wages	0.050	0.002	0.164
	(0.035)	(0.026)	(0.028)
Numerator Wald-TC	0.024	-0.012	0.103
	(0.026)	(0.021)	(0.028)
Numerator Wald-CIC	0.023	-0.009	0.099
	(0.027)	(0.021)	(0.028)
N	14452	19938	22339

Notes. This table reports placebo and true estimates comparing the evolution of education and wages from cohort -2 to 1 in our two groups of districts. Standard errors are clustered at the district level. For the numerator of the Wald-CIC, clustered standard errors are obtained by block bootstrap.

6.2 Results

First, we compare the weighted average of Wald-DIDs in Duflo (2001) to a simple Wald-DID with our control groups. In Table 6, we estimate the same 2SLS regression as that reported in the first column and third line of Table 7 in Duflo (2001), and we obtain returns of 7.3%

per year of schooling.[15] Then, we estimate the Wald-DID with our groups and find returns of 15.9% per year of schooling. This coefficient is significantly different from the previous one (t-stat=-2.15), and it is also more precisely estimated: its standard error is 37% smaller, presumably because it relies on a much larger first stage. While the estimator in Duflo (2001) is only significant at the 10% level (t-stat=1.68),[16] our Wald-DID is significant at any conventional level. Note that the difference between these two estimators does not come from the fact they are estimated on different samples. Estimating Duflo's regression on our sample of 22,339 observations actually yields a smaller coefficient than her original estimate, which is still significantly different from ours. The difference between these two estimates could stem from the fact that districts where years of schooling increased less also have higher returns to education. This would bias downward the estimate in Duflo (2001), while our Wald-DID estimator does not rely on any treatment effect homogeneity assumption.

On the other hand, the validity of our Wald-DID still relies on Assumption 4O, which might not be plausible in this context. For instance, under Assumption 4O the wage gap between high-school graduates in cohort 0 and 1 should remain the same if they had only completed primary school. Had they only completed primary school, high school graduates of both cohorts would have joined the labor market earlier, and would have had more labor market experience at the time we compare their wages. The wage gap between the two cohorts might then have been lower, because returns to experience tend to be decreasing (see e.g. Mincer & Jovanovic, 1979).[17] The data lends some support to this hypothesis. In the control group, while high-school graduates in cohort 1 earn 54% less than their cohort 0 counterpart, the gap is only 20% for non-graduates, and the difference is significant (t-stat=-7.64). This difference could partly arise from selection effects: non-graduates differ from high school graduates, so the cohort gap among non-graduates might not be equal to the cohort gap we would have observed among graduates had they not graduated. Still, it seems unlikely that selection can fully account for this almost threefold difference.

Our Wald-TC and Wald-CIC estimators do not rely on Assumption 4O. They lie in-between the estimate in Duflo (2001) and our Wald-DID. They do not differ significantly from the coefficient in Duflo (2001), but this is partly because this coefficient is imprecisely estimated. Using the Wald-TC estimator, one can for instance reject that returns to education are lower than 6% at the 5% level. On the other hand, the Wald-TC and Wald-CIC significantly differ from the Wald-DID, with t-stats respectively equal to -3.52 and -3.66. The Wald-DID and

[15] Our coefficient differs very slightly from that of the author because we were not able to obtain exactly her sample of 31,061 observations.

[16] This point estimate was significant at the 5% level in the original paper. But once clustering standard errors at the district level, which has become standard practice in DID analysis since Bertrand et al. (2004), it loses some statistical significance.

[17] We follow Mincer & Jovanovic (1979) and estimate a mincerian regression of wages on education, education squared, age, and age squared in our data. We also find a significantly negative coefficient of age squared.

Wald-TC rely on different "common trends" assumptions between districts (Assumptions 3 and 5O). But challenging one while defending the other seems difficult as these two assumptions are substantively very close. On the other hand, the Wald-TC and Wald-CIC do not require that the wage gap between cohorts be constant across potential levels of education (Assumption 4O). As discussed in the previous paragraph, this assumption is not warranted in this context. We therefore choose the Wald-TC and Wald-CIC as our preferred estimators.[18]

Table 6: Returns to education

	Duflo (2001)	W_{DID}	W_{TC}	W_{CIC}	OLS
Returns to education	0.073	0.159	0.104	0.100	0.077
	(0.043)	(0.028)	(0.027)	(0.027)	(0.001)
N	30828	22339	22339	22339	30828

Notes. This table reports estimates of returns to schooling. Standard errors are clustered at the district level. For the Wald-TC and Wald-CIC, clustered standard errors are obtained by block bootstrap.

As shown in Theorem 4.4, the parameter we estimate is a weighted average of the effect of increasing years of education from $d-1$ to d, over all possible values of d. The weights w_d can be estimated. They are shown in Figure 1. Our parameter puts the most weight on the last years of primary school, on middle-school years, and on high-school years. Because in the treatment group the distribution of education in young cohorts does not dominate that in old cohorts, some weights are negative. But negative weights are fairly small, and sum up to -0.14. Therefore, failure of stochastic dominance is unlikely to drive our results.

[18] To estimate the numerator of the Wald-CIC, we do not estimate Q_d for each year of schooling. Instead, we group schooling into 5 categories (did not complete primary school, completed primary school, completed middle school, completed high school, completed college). Thus, we avoid estimating quantile-quantile transforms on a very small number of units. To be consistent, we also use this definition to estimate the numerator of the Wald-TC. Using years of schooling hardly changes our Wald-TC estimator.

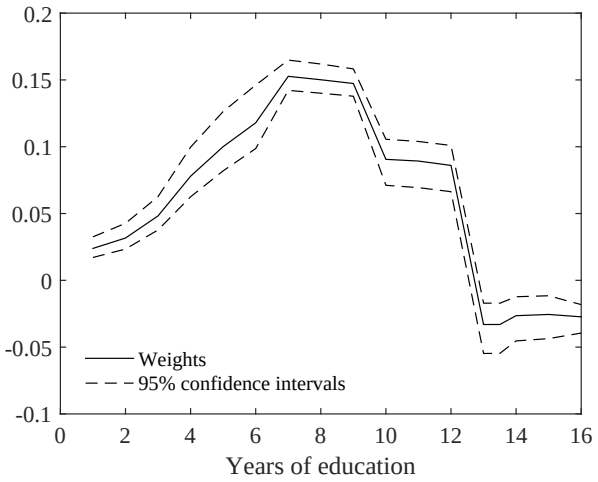

Figure 1: Weight received by each year of education.

6.3 Robustness checks

As a first robustness check, we investigate whether misclassifications of treatment districts as controls can bias our results. To do so, we construct our groups again using a more liberal criterion. Specifically, we assign a district to the control group if the p-value of the chi-squared test is greater than 0.6. If that p-value is lower than 0.6 and the average number of years of education increased in that district, we assign it to the treatment group. The control group we obtain this way is 30% smaller than the previous one, which increases the variance of our estimators. It also has a more stable distribution of education: a chi-squared test does not reject the assumption that this distribution is the same between the two cohorts. On the other hand, using this new control group leaves our estimates essentially unchanged: the Wald-DID, Wald-TC, and Wald-CIC are now respectively equal to 15.8, 9.8, and 9.6%. This suggests that the small changes in the distribution of education in our control group shown in Table 4 do not drive our results.

As a second robustness check, we investigate whether the statistical procedure we use to form our groups biases our estimates. Our method uses the same data twice, to form groups and to estimate returns to education. It therefore shares some similarities with the endogenous stratification methods studied in Abadie et al. (2013), which can produce finite sample biases. We conduct a simulation study to investigate the determinants of the bias. We find that finite sample bias is increasing with the correlation between the treatment and the unobserved determinants of the outcome,[19] decreasing with the size of the groups where the first stage chi-squared tests are conducted, and decreasing with the change of treatment intensity in the

[19] An important difference with the methods studied in Abadie et al. (2013) is that our method does not use the outcome but the treatment to construct groups. Therefore, our method produces biased estimates only if the treatment is strongly correlated with the unobserved determinants of the outcome. If treatment is

population. To detect potential biases, Abadie et al. (2013) suggest comparing the baseline estimator to a split-sample estimator where half of the sample is used to construct groups, while the other half is used to compute the estimator. Our simulations also suggest this is a good way to assess the seriousness of the problem. With DGPs for which our procedure generates little or no bias, the split-sample and baseline estimators are very close from each other; on the other hand, with DGPs for which our procedure generates more bias, the split-sample and baseline estimators are far away. Therefore, we re-estimate 200 times our Wald-DID, Wald-TC, and Wald-CIC estimators using a split-sample procedure. The average of the split-sample estimators are respectively 17.7%, 8.5%, and 8.0%. The three split-sample estimators are not significantly different and less than 20% away from the original estimators. Overall, endogenous stratification does not seem to be a strong concern in this application.

As a last robustness check, we investigate whether accounting for the sampling variance induced by our classification procedure would greatly affect our conclusions. Doing so is not straightforward. A natural idea is to use a two-step bootstrap where in a first step we bootstrap individuals within each cohort of each district and run our procedure to form our control and treatment groups, while in a second step we bootstrap districts and estimate the Wald-DID, the Wald-TC, and the Wald-CIC. In practice, this procedure does not work well. Under the null that the distribution of education did not change over time, one can show that the bootstrap statistics we use in our chi-squared tests do not have an approximate chi-squared distribution, but are approximately distributed as sums of squares of $\mathcal{N}(0,2)$ variables.[20] We therefore classify much fewer districts as controls than in the original sample. Dividing the bootstrap test statistics by two does not solve the problem, because the modified statistic then has a different distribution from that of the original statistic under the alternative hypothesis. Instead, we opt for a modified version of the two-step bootstrap: as in the original sample we classify 23% of districts as controls, in each bootstrap replication we classify the 23% of districts with the lowest chi-squared statistic as controls. The standard errors of our three estimators are now respectively equal to 0.044, 0.045, and 0.045. Thus, accounting for the sampling variance in our first step procedure seems to increase notably the standard errors of our estimators, but also leaves our main conclusions unchanged. For instance, our Wald-DID estimator would still be significantly different from the Wald-TC and Wald-CIC with these larger standard errors. However, proving that this procedure indeed reproduces well the distribution of our estimators goes beyond the scope of this paper and is left for future work.

This application differs from other applications of the fuzzy DID method in two important ways. First, it makes use of individual-level data. Many applications of the fuzzy DID method we found in our literature review directly use aggregate data at the county $\times$ year or state $\times$ year level. Second, the set of districts where education did not change between the two cohorts

exogenous or only weakly endogenous, it does not produce biases.

[20] Because districts are of finite size, the distribution of the test statistic is not exactly equal to its asymptotic distribution.

is not known to the analyst and needs to be estimated. In many applications of the fuzzy DID method the set of groups where treatment is stable is known to the analyst (examples include Draca et al., 2011, Field, 2007, or Gentzkow et al., 2011). In our supplementary material, we revisit Gentzkow et al. (2011) who use aggregate data and where the set of groups where treatment is stable is known. We show that the methods we propose in this paper can also be applied to this type of data, and that they can lead to substantially different conclusions from those reached by the authors using existing methods.

7 Conclusion

This paper studies treatment effects estimation in fuzzy DID designs. It makes the following contributions. First, we show that the Wald-DID is equal to a local average treatment effect (LATE) only if two strong assumptions are satisfied: treatment effects should be constant over time, and when treatment increases both in the treatment and in the control group treatment effects should be homogeneous in the two groups. Second, we propose two alternative estimators for the same LATE when the distribution of treatment is stable over time in the control group. Our first estimator is a natural generalization of DID to the fuzzy case. Our second estimator generalizes the changes-in-changes estimator introduced by Athey & Imbens (2006). Our estimators do not require that treatment effects be stable over time. Third, we show that under the same assumptions as those underlying our estimators, the same LATE can be bounded when the distribution of treatment changes over time in the control group.

When using the DID method with fuzzy groups, it is crucial to find a control group where treatment is stable over time to achieve point identification without imposing treatment effect homogeneity assumptions. In such instances, three estimators are available: the Wald-DID and our two alternative estimators. Using one or the other estimator can make a substantial difference, as we show in our application.

References

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

Abadie, A., Chingos, M. M. & West, M. R. (2013), Endogenous stratification in randomized experiments, Technical report, National Bureau of Economic Research.

Angrist, J., Chernozhukov, V. & Fernández-Val, I. (2006), 'Quantile regression under misspecification, with an application to the u.s. wage structure', *Econometrica* **74**(2), 539–563.

Angrist, J. D., Graddy, K. & Imbens, G. W. (2000), 'The interpretation of instrumental variables estimators in simultaneous equations models with an application to the demand for fish', *The Review of Economic Studies* **67**(3), 499–527.

Angrist, J. D. & Imbens, G. W. (1995), 'Two-stage least squares estimation of average causal effects in models with variable treatment intensity', *Journal of the American Statistical Association* **90**(430), pp. 431–442.

Athey, S. & Imbens, G. W. (2006), 'Identification and inference in nonlinear difference-in-differences models', *Econometrica* **74**(2), 431–497.

Baten, J., Bianchi, N. & Moser, P. (2014), 'Patents, competition, and innovation-evidence from compulsory licensing during wwi', *Available at SSRN 2417532* .

Bertrand, M., Duflo, E. & Mullainathan, S. (2004), 'How much should we trust differences-in-differences estimates?', *The Quarterly Journal of Economics* **119**(1), 249–275.

Blundell, R., Dias, M. C., Meghir, C. & Reenen, J. V. (2004), 'Evaluating the employment impact of a mandatory job search program', *Journal of the European Economic Association* **2**(4), 569–606.

Bonhomme, S. & Sauder, U. (2011), 'Recovering distributions in difference-in-differences models: A comparison of selective and comprehensive schooling', *Review of Economics and Statistics* **93**(2), 479–494.

Chernozhukov, V., Fernández-Val, I. & Galichon, A. (2010), 'Quantile and probability curves without crossing', *Econometrica* **78**(3), 1093–1125.

Chernozhukov, V., Fernández-Val, I., Hahn, J. & Newey, W. (2013), 'Average and quantile effects in nonseparable panel models', *Econometrica* **81**(2), 535–580.

Chernozhukov, V., Fernández-Val, I. & Melly, B. (2013), 'Inference on counterfactual distributions', *Econometrica* **81**(6), 2205–2268.

de Chaisemartin, C. (2013), A note on the assumptions underlying instrumented differences-in-differences., Working paper.

de Chaisemartin, C. & D'Haultfœuille, X. (2014), Fuzzy changes-in-changes, Technical report.

de Chaisemartin, C. & D'Haultfœuille, X. (2015), Supplement to "fuzzy differences-in-differences", Technical report.

D'Haultfœuille, X., Hoderlein, S. & Sasaki, Y. (2013), Nonlinear difference-in-differences in repeated cross sections with continuous treatments. CEMMAP Working Paper CWP40/13.

Draca, M., Machin, S. & Witt, R. (2011), 'Panic on the streets of london: Police, crime, and the july 2005 terror attacks', *The American Economic Review* pp. 2157–2181.

Duflo, E. (2001), 'Schooling and labor market consequences of school construction in indonesia: Evidence from an unusual policy experiment', *American Economic Review* **91**(4), 795–813.

Field, E. (2007), 'Entitled to work: Urban property rights and labor supply in Peru', *The Quarterly Journal of Economics* **122**(4), 1561–1602.

Frölich, M. (2007), 'Nonparametric iv estimation of local average treatment effects with covariates', *Journal of Econometrics* **139**(1), 35–75.

Gentzkow, M., Shapiro, J. M. & Sinkinson, M. (2011), 'The effect of newspaper entry and exit on electoral politics', *The American Economic Review* **101**(7), 2980.

Heckman, J. J. & Robb, R. (1985), 'Alternative methods for evaluating the impact of interventions: An overview', *Journal of econometrics* **30**(1), 239–267.

Horowitz, J. L. & Manski, C. F. (1995), 'Identification and robustness with contaminated and corrupted data', *Econometrica* **63**(2), 281–302.

Imbens, G. W. & Rubin, D. B. (1997), 'Estimating outcome distributions for compliers in instrumental variables models', *Review of Economic Studies* **64**(4), 555–574.

Lee, D. S. (2009), 'Training, wages, and sample selection: Estimating sharp bounds on treatment effects', *Review of Economic Studies* **76**(3), 1071–1102.

Manski, C. F. (1990), 'Nonparametric bounds on treatment effects', *American Economic Review* **80**(2), 319–23.

Melly, B. & Santangelo, G. (2015), The changes-in-changes model with covariates.

Mincer, J. & Jovanovic, B. (1979), 'Labor mobility and wages'.

Newey, W. K. (1994), 'The asymptotic variance of semiparametric estimators', *Econometrica* **62**(6), pp. 1349–1382.

van der Vaart, A. W. (2000), *Asymptotic Statistics*, Cambridge Series in Statistical and Probabilistic Mathematics.

van der Vaart, A. W. & Wellner, J. A. (1996), *Weak convergence and Empirical Processes*, Springer.

Vytlacil, E. (2002), 'Independence, monotonicity, and latent index models: An equivalence result', *Econometrica* **70**(1), 331–341.

A Main proofs

The lemmas prefixed by S are stated and proven in our supplementary material (see de Chaisemartin & D'Haultfœuille, 2015). For any $\Theta \subset \mathbb{R}^k$, let $\overset{\circ}{\Theta}$ denote its interior and let $\mathcal{C}^0(\Theta)$ and $\mathcal{C}^1(\Theta)$ denote respectively the set of continuous functions and the set of continuously differentiable functions with strictly positive derivative on Θ. We most often use these notations with $\Theta = \mathcal{S}(Y)$, in which cases we simply denote these sets by $\mathcal{C}^0$ and $\mathcal{C}^1$ respectively. Finally, for any $(d,g,t) \in \mathcal{S}(D)\times\mathcal{S}(G)\times\mathcal{S}(T)$, let $p_{gt} = P(G=g,T=t)$, $p_{dgt} = P(D=d,G=g,T=t)$, $p_{d|gt} = P(D_{gt}=d)$, and $F_{dgt} = F_{Y_{dgt}}$.

Theorem 3.1

Proof when $p_{1|01} \geq p_{1|00}$

Assume $p_{1|01} \geq p_{1|00}$. By Assumption 2, $p_{1|11} > p_{1|10}$. Therefore, the threshold model on D and Assumption 1 imply that

$$v_{g1} \leq v_{00}, \text{ for } g \in \{0,1\}. \tag{12}$$

Then, it follows from Model (1) and Assumption 1 that

$$\begin{aligned} p_{1|g1} - p_{1|g0} &= P(V \geq v_{g1}|T=1,G=g) - P(V \geq v_{00}|T=0,G=g) \\ &= P(V \in [v_{g1}, v_{00})|G=g). \end{aligned} \tag{13}$$

For any $g \in \{0,1\}$,

$$\begin{aligned} & E(Y_{g1}) - E(Y_{g0}) \\ = \; & E(h_D(U_D,1)|G=g,T=1) - E(h_D(U_D,0)|G=g,T=0) \\ = \; & E(h_1(U_1,1)|G=g,V \geq v_{g1})P(V \geq v_{g1}|G=g) + E(h_0(U_0,1)|G=g,V < v_{g1})P(V < v_{g1}|G=g) \\ - \; & E(h_1(U_1,0)|G=g,V \geq v_{00})P(V \geq v_{00}|G=g) - E(h_0(U_0,0)|G=g,V < v_{00})P(V < v_{00}|G=g) \\ = \; & E(h_1(U_1,1) - h_0(U_0,1)|G=g,V \in [v_{g1},v_{00}))P(V \in [v_{g1},v_{00})|G=g) \\ + \; & E(h_1(U_1,1) - h_1(U_1,0)|G=g,V \geq v_{00})P(V \geq v_{00}|G=g) \\ + \; & E(h_0(U_0,1) - h_0(U_0,0)|G=g,V < v_{00})P(V < v_{00}|G=g) \\ = \; & E(Y_{g1}(1) - Y_{g1}(0)|V \in [v_{g1},v_{00}))P(V \in [v_{g1},v_{00})|G=g) \\ + \; & E(h_1(U_1,1) - h_1(U_1,0)|G=g,V \geq v_{00})P(V \geq v_{00}|G=g) \\ + \; & E(h_0(U_0,1) - h_0(U_0,0)|G=g,V < v_{00})P(V < v_{00}|G=g) \\ = \; & E(Y_{g1}(1) - Y_{g1}(0)|V \in [v_{g1},v_{00}))P(V \in [v_{g1},v_{00})|G=g) \\ + \; & E(h_0(U_0,1) - h_0(U_0,0)|G=g). \end{aligned} \tag{14}$$

The first, second, third, fourth, and fifth equalities respectively follow from Model (1), Model (1) and Assumption 1, Equation (12), Model (1) and Assumption 1, and Assumption 4.

Combining Equation (14) and Assumption 3 imply that

$$\begin{aligned} DID_Y =& E(Y_{11}(1) - Y_{11}(0)|S_1)P(S_1|G=1) \\ & - E(Y_{01}(1) - Y_{01}(0)|S_0)P(S_0|G=0). \end{aligned}$$

Dividing each side by DID_D and using Equation (13) yields the result.

Proof when $p_{1|01} < p_{1|00}$

Assume $p_{1|01} < p_{1|00}$. Equation (14) still holds for $g = 1$, but not for $g = 0$ because $v_{00} < v_{01}$. On the other hand, a reasoning similar to that we used to derive Equations (13) yields

$$p_{1|00} - p_{1|01} \quad = \quad P(S_0|G=0). \tag{15}$$

Moreover,

$$\begin{aligned} & E(Y_{01}) - E(Y_{00}) \\ = \; & E(h_1(U_1,1)|G=0, V \geq v_{01})P(V \geq v_{01}|G=0) + E(h_0(U_0,1)|G=0, V < v_{01})P(V < v_{01}|G=0) \\ - \; & E(h_1(U_1,0)|G=0, V \geq v_{00})P(V \geq v_{00}|G=0) - E(h_0(U_0,0)|G=0, V < v_{00})P(V < v_{00}|G=0) \\ = \; & -E(h_1(U_1,1) - h_0(U_0,1)|G=0, V \in [v_{00}, v_{01}))P(V \in [v_{00}, v_{01})|G=0) \\ + \; & E(h_1(U_1,1) - h_1(U_1,0)|G=0, V \geq v_{00})P(V \geq v_{00}|G=0) \\ + \; & E(h_0(U_0,1) - h_0(U_0,0)|G=0, V < v_{00})P(V < v_{00}|G=0) \\ = \; & -E(Y_{01}(1) - Y_{01}(0)|V \in [v_{00}, v_{01}))P(V \in [v_{00}, v_{01})|G=0) \\ + \; & E(h_0(U_0,1) - h_0(U_0,0)|G=0). \end{aligned} \tag{16}$$

The first, second, and third equalities respectively follow from Model (1) and Assumption 1, $v_{00} < v_{01}$, Model (1) and Assumption 1 and 4. Taking the difference between Equation (14) with $g = 1$ and Equation (16) yields

$$\begin{aligned} DID_Y =& E(Y_{11}(1) - Y_{11}(0)|S_1)P(S_1|G=1) \\ & + E(Y_{01}(1) - Y_{01}(0)|S_0)P(S_0|G=0). \end{aligned}$$

Dividing each side of the previous display by DID_D and using Equations (13) and (15) yields the result. □

Theorem 3.2

Proof of 1

Following the same steps as those used to reach the last but one equality in Equation (14), we obtain

$$\begin{aligned} & E(Y_{11}) - E(Y_{10}) \\ = \; & E(Y_{11}(1) - Y_{11}(0)|S_1)P(S_1|G=1) \\ + \; & E(h_1(U_1,1) - h_1(U_1,0)|G=1, V \geq v_{00})P(V \geq v_{00}|G=1) \\ + \; & E(h_0(U_0,1) - h_0(U_0,0)|G=1, V < v_{00})P(V < v_{00}|G=1). \end{aligned} \tag{17}$$

Then,

$$\begin{aligned}\delta_1 &= E(Y_{101}) - E(Y_{100}) \\ &= E(h_1(U_1,1)|G=0, V \geq v_{00}) - E(h_1(U_1,0)|G=0, V \geq v_{01}) \\ &= E(h_1(U_1,1) - h_1(U_1,0)|G=0, V \geq v_{00}). \end{aligned} \tag{18}$$

The second equality follows from Model (1) and Assumption 1. The third one follows from the fact that $p_{1|01} = p_{1|00}$ combined with Assumption 1 implies that $\{G=0, V \leq v_{01}\} = \{G=0, V \leq v_{00}\}$.

Similarly,

$$\delta_0 = E(h_0(U_0,1) - h_0(U_0,0)|G=0, V < v_{00}). \tag{19}$$

Finally, the result follows combining Equations (17), (18), (19), and Assumption 5, once noted that $p_{1|10} = P(V \geq v_{00}|G=1)$ and $P(S_1|G=1) = p_{1|11} - p_{1|10}$.

Proof of 2

We only prove that $\underline{W}_{TC}$ is a lower bound when $\lambda_{00} > 1$. The proofs for the upper bound and when $\lambda_{00} < 1$ are symmetric.

We have

$$\begin{aligned} & E(Y_{11}(1) - Y_{11}(0)|S_1)P(S_1|G=1) \\ = \; & E(Y_{11}) - E(Y_{10}) \\ - \; & E(h_1(U_1,1) - h_1(U_1,0)|G=1, V \geq v_{00})P(V \geq v_{00}|G=1) \\ - \; & E(h_0(U_0,1) - h_0(U_0,0)|G=1, V < v_{00})P(V < v_{00}|G=1) \\ = \; & E(Y_{11}) - E(Y_{10}) \\ - \; & E(h_1(U_1,1) - h_1(U_1,0)|G=0, V \geq v_{00})P(V \geq v_{00}|G=1) \\ - \; & E(h_0(U_0,1) - h_0(U_0,0)|G=0, V < v_{00})P(V < v_{00}|G=1) \\ = \; & E(Y_{11}) - E(Y_{10}) \\ - \; & (E(Y_{01}(1)|V \geq v_{00}) - E(Y_{100}))p_{1|10} \\ - \; & (E(Y_{01}(0)|V < v_{00}) - E(Y_{000}))p_{0|10}. \end{aligned}$$

The first, second, and third equalities respectively follow from Equation (17), Assumption 5, and Model (1) combined with Assumption 1.

It follows from the last display that the proof will be complete if we can show that $\overline{\delta}_1$ and $\overline{\delta}_0$ are respectively upper bounds for $E(Y_{01}(1)|V \geq v_{00}) - E(Y_{100})$ and $E(Y_{01}(0)|V < v_{00}) - E(Y_{000})$.

When $\lambda_{00} > 1$, it follows from Model (1) and Assumption 1 that $v_{00} < v_{01}$. Then, we have

$$\begin{aligned} P(V \geq v_{01}|G=0,T=1,V \geq v_{00}) &= \frac{P(V \geq v_{01}|G=0,T=1)}{P(V \geq v_{00}|G=0,T=1)} \\ &= \frac{P(V \geq v_{01}|G=0,T=1)}{P(V \geq v_{00}|G=0,T=0)} \\ &= \frac{p_{1|01}}{p_{1|00}} \\ &= \lambda_{01}, \end{aligned} \tag{20}$$

where the second equality follows from Assumption 1. Therefore,

$$\begin{aligned} E(Y_{01}(1)|V \geq v_{00}) &= \lambda_{01}E(Y_{01}(1)|V \geq v_{01}) + (1-\lambda_{01})E(Y_{01}(1)|V \in S_0) \\ &\leq \lambda_{01}E(Y_{101}) + (1-\lambda_{01})\overline{y} = \int y d\underline{F}_{101}(y). \end{aligned} \tag{21}$$

This proves that $\overline{\delta}_1$ is an upper bound for $E(Y_{01}(1)|V \geq v_{00}) - E(Y_{100})$.

Similarly,

$$P(V < v_{00}|G=0,T=1,V < v_{01}) = 1/\lambda_{00},$$

and

$$E(Y_{001}) = 1/\lambda_{00}E(Y_{01}(0)|V < v_{00}) + (1-1/\lambda_{00})\,E(Y_{01}(0)|V \in S_0).$$

Following Horowitz & Manski (1995), the last display implies that

$$E(Y_{01}(0)|V < v_{00}) \leq \int y d\underline{F}_{001}(y).$$

This proves that $\overline{\delta}_0$ is an upper bound for $E(Y_{01}(0)|V < v_{00}) - E(Y_{000})$. □

Lemma 3.1

We only prove the formula for $d = 0$, the reasoning being similar for $d = 1$.

Using the same steps as those used to prove Equations (20) and (21) , one can show that

$$P(S_1|G=1,T=1,V < v_{00}) = \frac{p_{0|10} - p_{0|11}}{p_{0|10}}$$

and

$$F_{Y_{11}(0)|V<v_{00}}(y) = \frac{p_{0|10} - p_{0|11}}{p_{0|10}} F_{Y_{11}(0)|S_1}(y) + \frac{p_{0|11}}{p_{0|10}} F_{011}(y).$$

Therefore,

$$F_{Y_{11}(0)|S_1}(y) = \frac{p_{0|10}F_{Y_{11}(0)|V<v_{00}}(y) - p_{0|11}F_{011}(y)}{p_{0|10} - p_{0|11}}. \tag{22}$$

Then, we show that for all $y \in \mathcal{S}(Y_{11}(0)|V < v_{00})$,

$$F_{Y_{11}(0)|V<v_{00}} = F_{010} \circ F_{000}^{-1} \circ F_{Y_{01}(0)|V<v_{00}}. \tag{23}$$

Assumption 1 implies that $U_0 \perp\!\!\!\perp T|G, V < v_{00}$. As a result, for all $(g,t) \in \{0,1\}^2$,

$$\begin{aligned} F_{Y_{gt}(0)|V<v_{00}}(y) &= P(h_0(U_0,t) \leq y|G=g, T=t, V<v_{00}) \\ &= P(U_0 \leq h_0^{-1}(y,t)|G=g, V<v_{00}) \\ &= F_{U_0|G=g,V<v_{00}}(h_0^{-1}(y,t)). \end{aligned}$$

The second point of Assumption 7 combined with Assumptions 1 and 6 implies that $F_{U_0|G=g,V<v_{00}}$ is strictly increasing. Hence, its inverse exists and for all $q \in (0,1)$,

$$F^{-1}_{Y_{gt}(0)|V<v_{00}}(q) = h_0\left(F^{-1}_{U_0|G=g,V<v_{00}}(q), t\right).$$

This implies that for all $y \in \mathcal{S}(Y_{g1}(0)|V < v_{00})$,

$$F^{-1}_{Y_{g0}(0)|V<v_{00}} \circ F_{Y_{g1}(0)|V<v_{00}}(y) = h_0(h_0^{-1}(y,1),0). \tag{24}$$

By Assumption 7, we have

$$\begin{aligned} &\mathcal{S}(Y_{010}) = \mathcal{S}(Y_{000}) \\ &\Rightarrow \mathcal{S}(Y_{10}(0)|V<v_{00}) = \mathcal{S}(Y_{00}(0)|V<v_{00}) \\ &\Rightarrow \mathcal{S}(h_0(U_0,0)|V<v_{00}, G=1, T=0) = \mathcal{S}(h_0(U_0,0)|V<v_{00}, G=0, T=0) \\ &\Rightarrow \mathcal{S}(U_0|V<v_{00}, G=1) = \mathcal{S}(U_0|V<v_{00}, G=0) \\ &\Rightarrow \mathcal{S}(h_0(U_0,1)|V<v_{00}, G=1, T=1) = \mathcal{S}(h_0(U_0,1)|V<v_{00}, G=0, T=1) \\ &\Rightarrow \mathcal{S}(Y_{11}(0)|V<v_{00}) = \mathcal{S}(Y_{01}(0)|V<v_{00}), \end{aligned}$$

where the third and fourth implications are obtained combining Assumptions 1 and 6. Once combined with Equation (24), the previous display implies that for all $y \in \mathcal{S}(Y_{11}(0)|V < v_{00})$,

$$F^{-1}_{Y_{10}(0)|V<v_{00}} \circ F_{Y_{11}(0)|V<v_{00}}(y) = F^{-1}_{Y_{00}(0)|V<v_{00}} \circ F_{Y_{01}(0)|V<v_{00}}(y).$$

This proves Equation (23), because $\{V < v_{00}, G = g, T = 0\} = \{D = 0, G = g, T = 0\}$.

Finally, we show that

$$F_{Y_{01}(0)|V<v_{00}}(y) = \lambda_{00} F_{001}(y) + (1-\lambda_{00}) F_{Y_{01}(0)|S_0}(y). \tag{25}$$

Suppose first that $\lambda_{00} \leq 1$. Then, $v_{01} \leq v_{00}$ and $S_0 = \{V \in [v_{01}, v_{00}), G = 0\}$. Moreover, reasoning as for $P(S_1|G=1, V<v_{00})$, we get

$$\begin{aligned} \lambda_{00} &= \frac{P(V<v_{01}|G=0)}{P(V<v_{00}|G=0)} = P(V<v_{01}|G=0, V<v_{00}) \\ F_{Y_{01}(0)|V<v_{00}}(y) &= \lambda_{00} F_{001}(y) + (1-\lambda_{00}) F_{Y_{01}(0)|S_0}(y). \end{aligned}$$

If $\lambda_{00} > 1$, $v_{01} > v_{00}$ and $S_0 = \{V \in [v_{00}, v_{01}), G = 0\}$. We then have

$$1/\lambda_{00} = P(V < v_{00} | G = 0, V < v_{01})$$
$$F_{001}(y) = 1/\lambda_{00} F_{Y_{01}(0)|V<v_{00}}(y) + (1 - 1/\lambda_{00})\, F_{Y_{01}(0)|S_0}(y),$$

so Equation (25) is also satisfied.

The lemma follows by combining (22), (23) and (25). □

Theorem 3.3

Proof of 1

The proof follows from Lemma 3.1: $\lambda_{00} = \lambda_{01} = 1$ when $p_{d|00} = p_{d|01} > 0$.

Proof of 2

Construction of the bounds.

We only establish the validity of the bounds for $F_{Y_{11}(0)|S_1}(y)$. The reasoning is similar for $F_{Y_{11}(1)|S_1}(y)$. Bounds for Δ and τ_q directly follow from those for the cdfs.

We start considering the case where $\lambda_{00} < 1$. We first show that in such instances, $0 \leq T_0, G_0(T_0), C_0(T_0) \leq 1$ if and only if

$$\underline{T_0} \leq T_0 \leq \overline{T_0}. \tag{26}$$

$G_0(T_0)$ is included between 0 and 1 if and only if

$$\frac{-\lambda_{00} F_{001}}{1 - \lambda_{00}} \leq T_0 \leq \frac{1 - \lambda_{00} F_{001}}{1 - \lambda_{00}},$$

while $C_0(T_0)$ is included between 0 and 1 if and only if

$$\frac{H_0^{-1}(\lambda_{10} F_{011}) - \lambda_{00} F_{001}}{1 - \lambda_{00}} \leq T_0 \leq \frac{H_0^{-1}(\lambda_{10} F_{011} + (1 - \lambda_{10})) - \lambda_{00} F_{001}}{1 - \lambda_{00}}.$$

Since $-\lambda_{00} F_{001}/(1 - \lambda_{00}) \leq 0$ and $(1 - \lambda_{00} F_{001})/(1 - \lambda_{00}) \geq 1$, T_0, $G_0(T_0)$ and $C_0(T_0)$ are all included between 0 and 1 if and only if

$$M_0\left(\frac{H_0^{-1}(\lambda_{10} F_{011}) - \lambda_{00} F_{001}}{1 - \lambda_{00}}\right) \leq T_0 \leq m_1\left(\frac{H_0^{-1}(\lambda_{10} F_{011} + (1 - \lambda_{10})) - \lambda_{00} F_{001}}{1 - \lambda_{00}}\right). \tag{27}$$

Composing each term of these inequalities by $M_0(.)$ and then by $m_1(.)$ yields Equation (26), since $M_0(T_0) = m_1(T_0) = T_0$ and $M_0 \circ m_1 = m_1 \circ M_0$.

Now, when $\lambda_{00} < 1$, $G_0(T_0)$ is increasing in T_0, so $C_0(T_0)$ as well is increasing in T_0. Combining this with (26) implies that for every y',

$$C_0(\underline{T_0})(y') \leq C_0(T_0)(y') \leq C_0(\overline{T_0})(y'). \tag{28}$$

Because $C_0(T_0)(y)$ is a cdf,

$$C_0(T_0)(y) = \inf_{y' \geq y} C_0(T_0)(y') \leq \inf_{y' \geq y} C_0(\overline{T_0})(y') = \overline{F}_{CIC,0}(y).$$

This proves the result for the upper bound. The result for the lower bound follows similarly.

Let us now turn to the case where $\lambda_{00} > 1$. Using the same reasoning as above, we get that $G_0(T_0)$ and $C_0(T_0)$ are included between 0 and 1 if and only if

$$\begin{aligned} \frac{\lambda_{00}F_{001} - 1}{\lambda_{00} - 1} &\leq T_0 \leq \frac{\lambda_{00}F_{001}}{\lambda_{00} - 1}, \\ \frac{\lambda_{00}F_{001} - H_0^{-1}(\lambda_{10}F_{011} + (1 - \lambda_{10}))}{\lambda_{00} - 1} &\leq T_0 \leq \frac{\lambda_{00}F_{001} - H_0^{-1}(\lambda_{10}F_{011})}{\lambda_{00} - 1}. \end{aligned}$$

The inequalities in the first line are not binding since they are implied by those on the second line. Thus, we also get (27). Hence, $0 \leq T_0, G_0(T_0), C_0(T_0) \leq 1$ if and only if

$$\overline{T_0} \leq T_0 \leq \underline{T_0}. \tag{29}$$

Besides, when $\lambda_{00} > 1$, $G_0(T_0)$ is decreasing in T_0, so $C_0(T_0)$ is also decreasing in T_0. Combining this with Equation (29) implies that for every y, Equation (28) holds as well. This proves the result.

Sketch of the proof of sharpness.

The full proof is in the supplementary material (see de Chaisemartin & D'Haultfœuille, 2015). We only consider the sharpness of $\underline{F}_{CIC,0}$, the reasoning being similar for the upper bound. The proof is also similar and actually simpler for $d = 1$. The corresponding bounds are proper cdf, so we do not have to consider converging sequences of cdf as we do in case b) below.

a. $\lambda_{00} > 1$**.** We show that if Assumptions 7-9 hold, then $\underline{F}_{CIC,0}$ is sharp. For that purpose, we construct $\widetilde{h}_0, \widetilde{U}_0, \widetilde{V}$ such that:

(i) $Y = \widetilde{h}_0(\widetilde{U}_0, T)$ when $D = 0$ and $D = 1\{\widetilde{V} \geq v_{GT}\}$;

(ii) $(\widetilde{U}_0, \widetilde{V}) \perp\!\!\!\perp T|G$;

(iii) $\widetilde{h}_0(., t)$ is strictly increasing for $t \in \{0, 1\}$;

(iv) $F_{\widetilde{h}_0(\widetilde{U}_0,1)|G=0,T=1,\widetilde{V}\in[v_{00},v_{01})} = \underline{T}_0$.

(i) ensures that Model (1) is satisfied on the observed data. Because we can always define $\widetilde{Y}(0)$ as $\widetilde{h}_0(\widetilde{U}_0, T)$ when $D = 1$ without contradicting the data and the model, (i) is actually sufficient for Model (1) to hold globally, not only on the observed data. (ii) and (iii) ensure that Assumptions 1 and 6 hold. Finally, (iv) ensures that the DGP corresponding to $(\widetilde{h}_0, \widetilde{U}_0, \widetilde{V})$ rationalizes the bound.

The construction of $\widetilde{h}_0$, $\widetilde{U}_0$, and $\widetilde{V}$ is long, so its presentation is deferred to the supplementary material.

b. $\lambda_{00} < 1$. The idea is similar as in the previous case. A difference, however, is that when $\lambda_{00} < 1$, $\underline{T}_0$ is not a proper cdf, but a defective one, since $\lim_{y \to \overline{y}} \underline{T}_0(y) < 1$. As a result, we cannot define a DGP such that $\widetilde{T}_0 = \underline{T}_0$, However, by Lemma S2, there exists a sequence $(\underline{T}_0^k)_k$ of cdf such that $\underline{T}_0^k \to \underline{T}_0$, $G_0(\underline{T}_0^k)$ is an increasing bijection from $\mathcal{S}(Y)$ to $(0,1)$ and $C_0(\underline{T}_0^k)$ is increasing and onto $(0,1)$. We can then construct a sequence of DGP $(\widetilde{h}_0^k(.,0), \widetilde{h}_0^k(.,1), \widetilde{U}_0^k, \widetilde{V}^k)$ such that Points (i) to (iii) listed above hold for every k, and such that $\widetilde{T}_0^k = \underline{T}_0^k$. Since $\underline{T}_0^k(y)$ converges to $\underline{T}_0(y)$ for every y in $\mathcal{S}(\overset{\circ}{Y})$, we thus define a sequence of DGP such that $\widetilde{T}_0^k$ can be arbitrarily close to $\underline{T}_0$ on $\mathcal{S}(\overset{\circ}{Y})$ for sufficiently large k. Since $C_0(.)$ is continuous, this proves that $\underline{F}_{CIC,0}$ is sharp on $\mathcal{S}(\overset{\circ}{Y})$. This construction is long, so its exposition is deferred to the supplementary material. □

Theorem 3.4

Proof of 1

$p_{1|00} = p_{1|10}$ implies that $W_{DID} = W_{TC}$. Therefore, the proof will be complete if we can show that $W_{DID} = E(Y_{11}(1) - Y_{11}(0)|D = 1)$. On that purpose, notice that the outcome Equation of Model (1), $U_0 \perp\!\!\!\perp T|G$, and Assumption 3 imply that

$$E(Y_{11}(0)) - E(Y_{10}(0)) - (E(Y_{01}(0)) - E(Y_{00}(0))) = 0. \tag{30}$$

Then,

$$\begin{aligned} DID_Y &= E(Y_{11}) - E(Y_{10}) - (E(Y_{01}) - E(Y_{00})) \\ &= p_{1|11} E(Y_{11}(1) - Y_{11}(0)|D = 1) + E(Y_{11}(0)) - E(Y_{10}(0)) - (E(Y_{01}(0)) - E(Y_{00}(0))) \\ &= p_{1|11} E(Y_{11}(1) - Y_{11}(0)|D = 1). \end{aligned}$$

The second equality follows from $p_{1|00} = p_{1|01} = p_{1|10} = 0$, the third from Equation (30). This completes the proof once noted that $DID_D = p_{1|11}$.

Proof of 2

As $p_{1|10} = 0$, the numerator of W_{CIC} is $E(Y_{11}) - E(Q_0(Y_{10}))$. It is easy to see that the proof will be complete if we can show that $E(Q_0(Y_{10})) = E(Y_{11}(0))$. As $p_{1|00} = p_{1|01} = 0$, Q_0 is the quantile-quantile transform of the outcome in the entire control group, so $E(Q_0(Y_{10}))$ is the same estimand as that considered in Equation (16) in Athey & Imbens (2006). The outcome equation of Model (1), $U_0 \perp\!\!\!\perp T|G$, and Assumptions 6 and 7 ensure that the assumptions of their Theorem 3.1 hold. Therefore, $E(Q_0(Y_{10})) = E(Y_{11}(0))$ □

Theorem 3.5

Assume that $p_{1|00} = p_{1|01} = 1$ (the proof is symmetric when $p_{1|00} = p_{1|01} = 0$). For $F_{Y_{11}(1)|S_1}(y)$, the proof directly follows from Lemma 3.1. For $F_{Y_{11}(0)|S_1}(y)$, one can follow similar steps as those used to establish Equation (24) and show that for all $y \in \mathcal{S}(Y)$,

$$F^{-1}_{Y_{00}(1)|V \geq v_{00}} \circ F_{Y_{01}(1)|V \geq v_{00}}(y) = h_1(h_1^{-1}(y,1),0). \tag{31}$$

Equations (24) and (31), Assumption 10, and $p_{1|00} = p_{1|01} = 1$ imply that for all $y \in \mathcal{S}(Y)$,

$$F_{Y_{11}(0)|V < v_{00}}(y) \quad = \quad F_{010} \circ F^{-1}_{100} \circ F_{101}(y). \tag{32}$$

Combining Equations (22) and (32) yields the result □

Theorem 4.1

We start proving the first statement. Under the assumptions of the theorem, Assumptions 1-4 are satisfied for the treatment and control groups $G^*_t = 1$ and $G^*_t = 0$ between dates $t-1$ and t. For instance, the fact that $(U_d, V) \perp\!\!\!\perp T | G^*_t = 0$ follows from the fact that $G \perp\!\!\!\perp T$ and $(U_d, V) \perp\!\!\!\perp T | G = g$ for every $g \in \mathcal{G}_{st}$. Moreover, for every $t \geq 1$ and for every $g \in \mathcal{G}_{st}$, $E(D_{gt}) = E(D_{gt-1})$, thus implying that $E(D|G^*_t = 0, T = t) = E(D|G^*_t = 0, T = t-1)$. Therefore, it follows from Theorem 3.1 that

$$W^*_{DID}(1,0,t) = E(Y(1) - Y(0)|S_t, G^*_t = 1, T = t). \tag{33}$$

Similarly, one can show that

$$W^*_{DID}(-1,0,t) = E(Y(1) - Y(0)|S_t, G^*_t = -1, T = t). \tag{34}$$

Then, $G \perp\!\!\!\perp T$ implies that

$$\begin{aligned} DID^*_D(1,0,t)P(G^*_t = 1) &= (E(D|G^*_t = 1, T = t) - E(D|G^*_t = 1, T = t-1))P(G^*_t = 1) \\ &= P(S_t|G^*_t = 1)P(G^*_t = 1) \\ &= P(S_t, G^*_t = 1). \end{aligned}$$

Similarly, one can show that

$$DID^*_D(0,-1,t)P(G^*_t = -1) \quad = \quad P(S_t, G^*_t = -1).$$

Therefore, it follows from the two previous displays that

$$DID^*_D(1,0,t)P(G^*_t = 1) + DID^*_D(0,-1,t)P(G^*_t = -1) = P(S_t) \tag{35}$$

and

$$\begin{aligned} & \frac{DID^*_D(1,0,t)P(G^*_t = 1)}{DID^*_D(1,0,t)P(G^*_t = 1) + DID^*_D(0,-1,t)P(G^*_t = -1)} \\ = \quad & P(G^*_t = 1|S_t). \end{aligned} \tag{36}$$

The result follows combining Equations (33), (34), (35), and (36), once noted that Assumption 1 and $G \perp\!\!\!\perp T$ imply that $P(G_t^* = 1|S_t) = P(G_t^* = 1|S_t, T = t)$ and $P(G_t^* = -1|S_t) = P(G_t^* = -1|S_t, T = t)$.

The proofs of the second and third statements follow from similar arguments. To prove the fourth statement, it suffices to notice that the first point of Assumption 11 implies that for every $g \in \{0, 1, ..., \overline{g}\}$ the sequence v_{gt} is monotonic in t. Therefore, for every $g \in \mathcal{S}(G)$ and $t \neq t' \in \{1, ..., \overline{t}\}^2$, $S_{gt} \cap S_{gt'} = \emptyset$. This in turn implies that $S_t \cap S_{t'} = \emptyset$. Combining this with the third point of Assumption 11 yields the result □

Theorem 4.2

The two results are straightforward extensions of the second point of Theorems 3.2 and 3.3, so their proof is omitted.

Theorem 4.3

We only prove the first result, the second and third results follow from similar arguments.

$W_{DID}(X) = \Delta(X)$ follows from the same steps as those used to prove Theorem 3.1. Then, $W_{DID}^X = \Delta$ follows after some algebra, once noted that

$$\begin{aligned} f_{X_{11}|S_1}(x) &= \frac{E(D_{11}|X = x) - E(D_{10}|X = x)}{E(D_{11}) - E(E(D_{10}|X)|G = 1, T = 1)} f_{X_{11}}(x) \\ &= \frac{DID_D(x)}{E[DID_D(X)|G = 1, T = 1]} f_{X_{11}}(x). \end{aligned}$$

The first equality follows from Model (10), Assumption 1X, and Bayes's law. The second follows from the fact that $E(D_{01}|X) - E(D_{00}|X) = 0$ almost surely. □

Proof of Theorem 4.4

We only prove the first statement, the second and third statements follow from similar arguments.

$D_{01} \sim D_{00}$ and $D_{11} \gtrsim D_{10}$ combined with Model (11) and Assumption 1 imply that

$$v_{01}^d = v_{00}^d, \text{ for every } d \in \{1, \overline{d}\} \tag{37}$$

$$v_{11}^d \leq v_{10}^d, \text{ for every } d \in \{1, \overline{d}\}. \tag{38}$$

Then, it follows from Model (11), Assumption 1 and Equation (38) that for every $d \in \{1, 2, ..., \overline{d}\}$,

$$\begin{aligned} P(D_{11} \geq d) - P(D_{10} \geq d) &= P(V \geq v_{g1}^d | T = 1, G = g) - P(V \geq v_{g0}^d | T = 0, G = g) \\ &= P(V \in [v_{g1}^d, v_{g0}^d) | G = g). \end{aligned} \tag{39}$$

Then, for every $g \in \{0,1\}$,

$$
\begin{aligned}
& E(Y_{g1}) - E(Y_{g0}) \\
= \; & E(h_D(U_D,1)|G=g,T=1) - E(h_D(U_D,0)|G=g,T=0) \\
= \; & \sum_{d=0}^{\overline{d}} E(h_d(U_d,1)|G=g,V \in [v_{g1}^d, v_{g1}^{d+1}))P(V \in [v_{g1}^d, v_{g1}^{d+1})|G=g) \\
- \; & \sum_{d=0}^{\overline{d}} E(h_d(U_d,0)|G=g,V \in [v_{g0}^d, v_{g0}^{d+1}))P(V \in [v_{g0}^d, v_{g0}^{d+1})|G=g) \\
= \; & \sum_{d=1}^{\overline{d}} E(h_d(U_d,1) - h_{d-1}(U_{d-1},1)|G=g,V \in [v_{g1}^d, v_{g0}^d))P(V \in [v_{g1}^d, v_{g0}^d)|G=g) \\
+ \; & \sum_{d=0}^{\overline{d}} E(h_d(U_d,1) - E(h_d(U_d,0)|G=g,V \in [v_{g0}^d, v_{g0}^{d+1}))P(V \in [v_{g0}^d, v_{g0}^{d+1})|G=g) \\
= \; & \sum_{d=1}^{\overline{d}} E(Y_{g1}(d) - Y_{g1}(d-1)|V \in [v_{g1}^d, v_{g0}^d))P(V \in [v_{g1}^d, v_{g0}^d)|G=g) \\
+ \; & E(h_0(U_0,1) - h_0(U_0,0)|G=g). \qquad (40)
\end{aligned}
$$

The first, second, third, and fourth, equalities respectively follow from Model (11), Model (11) and Assumption 1, Equations (37) and (38), and Model (11) combined with Assumptions 1 and 4O.

Combining Equation (40) with Equation (37) and Assumption 3 imply that

$$DID_Y = \sum_{d=1}^{\overline{d}} E(Y_{11}(d) - Y_{11}(d-1)|V \in [v_{11}^d, v_{10}^d))P(V \in [v_{11}^d, v_{10}^d)|G=1).$$

The result follows from Equation (39), after dividing each side of the previous display by DID_D □

Theorem 5.1

Proof of 1 and 2

Asymptotic normality is obvious by the central limit theorem and the delta method. Consistency of the bootstrap follows by consistency of the bootstrap for sample means (see, e.g., van der Vaart, 2000, Theorem 23.4) and the delta method for bootstrap (van der Vaart, 2000, Theorem 23.5). A convenient way to obtain the asymptotic variance is to use repeatedly the following argument. If

$$\sqrt{n}\left(\widehat{A} - A\right) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} a_i + o_P(1) \text{ and } \sqrt{n}\left(\widehat{B} - B\right) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} b_i + o_P(1),$$

then Lemma S3 ensures that

$$\sqrt{n}\left(\frac{\widehat{A}}{\widehat{B}}-\frac{A}{B}\right)=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{a_i-(A/B)b_i}{B}+o_P(1). \tag{41}$$

This implies for instance that

$$\sqrt{n}\left(\widehat{E}(Y_{11})-E(Y_{11})\right)=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{G_iT_i(Y_i-E(Y_{11}))}{p_{11}}+o_P(1),$$

and similarly for $\widehat{E}(D_{11})$. Applying repeatedly this argument, we obtain, after some algebra,

$$\sqrt{n}\left(\widehat{W}_{DID}-\Delta\right)=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi_{DID,i}+o_P(1),$$

where, omitting the index i, ψ_{DID} is defined by

$$\begin{aligned}\psi_{DID}=\frac{1}{DID_D}\Bigg[&\frac{GT(\varepsilon-E(\varepsilon_{11}))}{p_{11}}-\frac{G(1-T)(\varepsilon-E(\varepsilon_{10}))}{p_{10}}-\frac{(1-G)T(\varepsilon-E(\varepsilon_{01}))}{p_{01}}\\&+\frac{(1-G)(1-T)(\varepsilon-E(\varepsilon_{00}))}{p_{00}}\Bigg]\end{aligned} \tag{42}$$

and $\varepsilon=Y-\Delta D$. Similarly,

$$\sqrt{n}\left(\widehat{W}_{TC}-\Delta\right)=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi_{TC,i}+o_P(1),$$

where ψ_{TC} is defined by

$$\begin{aligned}\psi_{TC}=&\frac{1}{E(D_{11})-E(D_{10})}\Bigg\{\frac{GT(\varepsilon-E(\varepsilon_{11}))}{p_{11}}-\frac{G(1-T)(\varepsilon+(\delta_1-\delta_0)D-E(\varepsilon_{10}+(\delta_1-\delta_0)D_{10}))}{p_{10}}\\&-E(D_{10})D(1-G)\left[\frac{T(Y-E(Y_{101}))}{p_{101}}-\frac{(1-T)(Y-E(Y_{100}))}{p_{100}}\right]\\&-(1-E(D_{10}))(1-D)(1-G)\left[\frac{T(Y-E(Y_{001}))}{p_{001}}-\frac{(1-T)(Y-E(Y_{000}))}{p_{000}}\right]\Bigg\}.\end{aligned} \tag{43}$$

Proof of 3

We first show that $(\widehat{F}_{Y_{11}(0)|S_1},\widehat{F}_{Y_{11}(1)|S_1})$ tends to a continuous gaussian process. Let $\widetilde{\theta}=(F_{000},F_{001},...,F_{111},\lambda_{10},\lambda_{11})$. By Lemma S4, $\widehat{\widetilde{\theta}}=(\widehat{F}_{000},\widehat{F}_{001},...,\widehat{F}_{111},\widehat{\lambda}_{10},\widehat{\lambda}_{11})$ converges to a continuous gaussian process. Let

$$\pi_d:(F_{000},F_{001},...,F_{111},\lambda_{10},\lambda_{11})\mapsto(F_{d10},F_{d00},F_{d01},F_{d11},1,\lambda_{1d}),\quad d\in\{0,1\},$$

so that $(\widehat{F}_{Y_{11}(0)|S_1},\widehat{F}_{Y_{11}(1)|S_1})=\left(R_1\circ\pi_0(\widehat{\widetilde{\theta}}),R_1\circ\pi_1(\widehat{\widetilde{\theta}})\right)$, where R_1 is defined as in Lemma S5. π_d is Hadamard differentiable as a linear continuous map. Because $F_{d10},F_{d00},F_{d01},F_{d11}$ are continuously differentiable with strictly positive derivative by Assumption 13, $\lambda_{1d}>0$, and

$\lambda_{1d} \neq 1$ under Assumption 7, R_1 is also Hadamard differentiable at $(F_{d10}, F_{d00}, F_{d01}, F_{d11}, 1, \lambda_{1d})$ tangentially to $(\mathcal{C}^0)^4 \times \mathbb{R}^2$. By the functional delta method (see, e.g., van der Vaart & Wellner, 1996, Lemma 3.9.4), $(\widehat{F}_{Y_{11}(0)|S_1}, \widehat{F}_{Y_{11}(1)|S_1})$ tends to a continuous gaussian process.

Now, by integration by parts for Lebesgue-Stieljes integrals,

$$\Delta = \int_{\underline{y}}^{\overline{y}} F_{Y_{11}(0)|S_1}(y) - F_{Y_{11}(1)|S_1}(y) dy.$$

Moreover, the map $\varphi_1 : (F_1, F_2) \mapsto \int_{\mathcal{S}(Y)} (F_2(y) - F_1(y)) dy$, defined on the domain of bounded càdlàg functions, is linear. Because $\mathcal{S}(Y)$ is bounded by Assumption 13, φ_1 is also continuous with respect to the supremum norm. It is thus Hadamard differentiable. Because $\widehat{\Delta} = \varphi_1 \left(\widehat{F}_{Y_{11}(1)|S_1}, \widehat{F}_{Y_{11}(0)|S_1} \right)$, $\widehat{\Delta}$ is asymptotically normal by the functional delta method. The asymptotic normality of $\widehat{\tau}_q$ follows along similar lines. By Assumption 13, $F_{Y_{11}(d)|S_1}$ is differentiable with strictly positive derivative on its support. Thus, the map $(F_1, F_2) \mapsto F_2^{-1}(q) - F_1^{-1}(q)$ is Hadamard differentiable at $(F_{Y_{11}(0)|S_1}, F_{Y_{11}(1)|S_1})$ tangentially to the set of functions that are continuous at $(F^{-1}_{Y_{11}(0)|S_1}(q), F^{-1}_{Y_{11}(1)|S_1}(q))$ (see Lemma 21.3 in van der Vaart, 2000). By the functional delta method, $\widehat{\tau}_q$ is asymptotically normal.

The validity of the bootstrap follows along the same lines. By Lemma S4, the bootstrap is consistent for $\widehat{\widehat{\theta}}$. Because both the LATE and LQTE are Hadamard differentiable functions of $\widehat{\widehat{\theta}}$, as shown above, the result simply follows by the functional delta method for the bootstrap (see, e.g., van der Vaart, 2000, Theorem 23.9).

Finally, we compute the asymptotic variance of both estimators. The functional delta method also implies that both estimators are asymptotically linear. To compute their asymptotic variance, it suffices to provide their asymptotic linear approximation. For that purpose, let us first linearize $F_{Y_{11}(d)|S_1}(y)$, for all y. It follows from the proof of the first point of Lemma S5 that the mapping $\phi_1 : (F_1, F_2, F_3) \mapsto F_1 \circ F_2^{-1} \circ F_3$ is Hadamard differentiable at $(F_{d10}, F_{d00}, F_{d01})$, tangentially to $(\mathcal{C}^0)^3$. Moreover applying the chain rule, we obtain

$$d\phi_1(h_1, h_2, h_3) = h_1 \circ Q_d^{-1} + H_d' \circ F_{d01} \times \left[-h_2 \circ Q_d^{-1} + h_3 \right].$$

Applied to $(F_1, F_2, F_3) = (F_{d10}, F_{d00}, F_{d01})$, this and the functional delta method once more imply that

$$\sqrt{n} \left(\widehat{H}_d \circ \widehat{F}_{d01} - H_d \circ F_{d01} \right) = d\phi_1(h_{1n}, h_{2n}, h_{3n}) + o_P(1),$$

where the $o_P(1)$ is uniform over y and $h_{1n} = \sqrt{n}(\widehat{F}_{d10} - F_{d10})$. h_{2n} and h_{3n} are defined similarly. Furthermore, applying Lemma S3 yields, uniformly over y,

$$h_{1n}(y) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{\mathbb{1}\{D_i = d\} G_i (1 - T_i)(\mathbb{1}\{Y_i \leq y\} - F_{d10}(y))}{p_{d10}} + o_P(1).$$

A similar expression holds for h_{2n} and h_{3n}. Hence, by continuity of $d\phi_1$, we obtain, after some algebra,

$$\begin{aligned}
&\sqrt{n}\left(\widehat{H}_d \circ \widehat{F}_{d01}(y) - H_d \circ F_{d01}(y)\right) \\
=&\frac{1}{\sqrt{n}}\sum_{i=1}^{n} \mathbb{1}\{D_i = d\}\left\{\frac{G_i(1-T_i)(\mathbb{1}\{Q_d(Y_i) \leq y\} - H_d \circ F_{d01}(y))}{p_{d10}} + (1-G_i)H'_d \circ F_{d01}(y)\right. \\
&\left.\times\left[-\frac{(1-T_i)(\mathbb{1}\{Q_d(Y_i) \leq y\} - F_{d01}(y))}{p_{d00}} + \frac{T_i(\mathbb{1}\{Y_i \leq y\} - F_{d01}(y))}{p_{d01}}\right]\right\} + o_P(1),
\end{aligned}$$

which holds uniformly over y. Applying repeatedly Lemma S3, we then obtain, after some algebra,

$$\sqrt{n}\left(\widehat{F}_{Y_{11}(d)|S_1}(y) - F_{Y_{11}(d)|S_1}(y)\right) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Psi_{di}(y) + o_P(1),$$

where, omitting the index i,

$$\begin{aligned}
\Psi_d(y) = &\frac{1}{p_{d|11} - p_{d|10}}\left\{\frac{GT}{p_{11}}\left[\mathbb{1}\{D=d\}\mathbb{1}\{Y \leq y\} - p_{d|11}F_{d11}(y) - F_{Y_{11}(d)|S_1}(y)\left(\mathbb{1}\{D=d\} - p_{d|11}\right)\right]\right. \\
&+ \frac{G(1-T)}{p_{10}}\left[-\mathbb{1}\{D=d\}\left(\mathbb{1}\{Q_d(Y) \leq y\} - H_d \circ F_{d01}(y)\right) + \left(\mathbb{1}\{D=d\} - p_{d|10}\right)\left(F_{Y_{11}(d)|S_1}(y) - H_d \circ F_{d01}(y)\right)\right] \\
&\left.+p_{d|10}(1-G)\mathbb{1}\{D=d\}H'_d \circ F_{d01}(y)\left[\frac{(1-T)(\mathbb{1}\{Q_d(Y) \leq y\} - F_{d01}(y))}{p_{d00}} - \frac{T(\mathbb{1}\{Y \leq y\} - F_{d01}(y))}{p_{d01}}\right]\right\}.
\end{aligned}$$

By the functional delta method, this implies that we can also linearize $\widehat{W}_{CIC}$ and $\widehat{\tau}_q$. Moreover, we obtain by the chain rule the following influence functions:

$$\psi_{CIC} = \int \Psi_0(y) - \Psi_1(y)dy, \tag{44}$$

$$\psi_{q,CIC} = \left[\frac{\Psi_1}{f_{Y_{11}(1)|S_1}}\right] \circ F^{-1}_{Y_{11}(1)|S_1}(q) - \left[\frac{\Psi_0}{f_{Y_{11}(0)|S_1}}\right] \circ F^{-1}_{Y_{11}(0)|S_1}(q). \tag{45}$$

Theorem 5.2

Proof of 1

For any random variable R, let $m^R_{gt}(x) = E(R_{gt}|X=x)$. The estimator $\widehat{W}^X_{DID}$ can be written as $\widehat{W}^X_{DID} = \widehat{N}^X_{DID}/\widehat{D}^X_{DID}$, with

$$\begin{aligned}
\widehat{N}^X_{DID} &= \widehat{E}\left[Y_{11}\right] - \widehat{E}\left[\widehat{m}^Y_{10}(X_{11})\right] - \widehat{E}\left[\widehat{m}^Y_{01}(X_{11})\right] + \widehat{E}\left[\widehat{m}^Y_{00}(X_{11})\right] \\
\widehat{D}^X_{DID} &= \widehat{E}\left[D_{11}\right] - \widehat{E}\left[\widehat{m}^D_{10}(X_{11})\right] - \widehat{E}\left[\widehat{m}^D_{01}(X_{11})\right] + \widehat{E}\left[\widehat{m}^D_{00}(X_{11})\right].
\end{aligned}$$

The true parameter $\Delta = N^X_{DID}/D^X_{DID}$ can be decomposed similarly. We show below that the eight terms in the numerator $\widehat{N}^X_{DID}$ and in the denominator $\widehat{D}^X_{DID}$ can be linearized. We can then use, as in the previous proof, the formula for linearizing ratios.

Let us first consider $\widehat{E}\left[\widehat{E}(Y_{10}|X)|G=1,T=1\right]$. Assumption 14 ensures that we can apply Lemma S8 to $I=G\times T$, $J=G\times(1-T)$, $U=Y$ and $V=1$. As a result,

$$\begin{aligned}
&\sqrt{n}\left(\widehat{E}\left[\widehat{E}(Y_{10}|X)|G=1,T=1\right]-E\left[m_{10}^{Y}(X)|G=1,T=1\right]\right)\\
=&\frac{1}{\sqrt{n}p_{11}}\sum_{i=1}^{n}G_i\left[T_i\left(m_{10}^{Y}(X_i)-E\left[m_{10}^{Y}(X)|G=1,T=1\right]\right)+\frac{(1-T_i)E(GT|X_i)}{E(G(1-T)|X_i)}\left(Y_i-m_{10}^{Y}(X_i)\right)\right]+o_P(1).
\end{aligned}$$

Applying the same reasoning as above to the two other terms of $\widehat{N}_{DID}^{X}$, we obtain

$$\begin{aligned}
&\sqrt{n}\left(\widehat{N}_{DID}^{X}-N_{DID}^{X}\right)\\
=&\frac{1}{\sqrt{n}p_{11}}\sum_{i=1}^{n}G_iT_i(Y_i-m_{10}^{Y}(X_i)-m_{01}^{Y}(X_i)+m_{00}^{Y}(X_i)-N_{DID}^{X})-\frac{G_i(1-T_i)E(GT|X_i)}{E(G(1-T)|X_i)}\left(Y_i-m_{10}^{Y}(X_i)\right)\\
&+\frac{(1-G_i)T_iE(GT|X_i)}{E((1-G)T|X_i)}\left(Y_i-m_{01}^{Y}(X_i)\right)-\frac{(1-G_i)(1-T_i)E(GT|X_i)}{E(1-G)(1-T)|X_i)}\left(Y_i-m_{00}^{Y}(X_i)\right)+o_P(1).
\end{aligned}$$

Similarly, the denominator satisfies

$$\begin{aligned}
&\sqrt{n}\left(\widehat{D}_{DID}^{X}-D_{DID}^{X}\right)\\
=&\frac{1}{\sqrt{n}p_{11}}\sum_{i=1}^{n}\left\{G_iT_i(D_i-m_{10}^{D}(X_i)-m_{01}^{D}(X_i)+m_{00}^{D}(X_i)-D_{DID}^{X})-\frac{G_i(1-T_i)E(GT|X_i)}{E(G(1-T)|X_i)}\left(D_i-m_{10}^{D}(X_i)\right)\right.\\
&\left.+\frac{(1-G_i)T_iE(GT|X_i)}{E((1-G)T|X_i)}\left(D_i-m_{01}^{D}(X_i)\right)-\frac{(1-G_i)(1-T_i)E(GT|X_i)}{E((1-G)(1-T)|X_i)}\left(D_i-m_{00}^{D}(X_i)\right)+o_P(1).\right.
\end{aligned}$$

Combining these two results and (41), we finally obtain

$$\sqrt{n}\left(\widehat{W}_{DID}^{X}-\Delta\right)=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi_{DID,i}^{X}+o_P(1),$$

where, omitting the index i, ψ_{DID}^{X} is defined by

$$\begin{aligned}
\psi_{DID}^{X}=&\frac{1}{p_{11}D_{DID}^{X}}\left\{GT(\varepsilon-m_{10}^{\varepsilon}(X)-m_{01}^{\varepsilon}(X)+m_{00}^{\varepsilon}(X))-\left[\frac{G(1-T)E(GT|X)}{E(G(1-T)|X)}\left(\varepsilon-m_{10}^{\varepsilon}(X)\right)\right.\right.\\
&\left.\left.+\frac{(1-G)TE(GT|X)}{E((1-G)T|X)}\left(\varepsilon-m_{01}^{\varepsilon}(X)\right)-\frac{(1-G)(1-T)E(GT|X)}{E((1-G)(1-T)|X)}\left(\varepsilon-m_{00}^{\varepsilon}(X)\right)\right]\right\},
\end{aligned}\tag{46}$$

and $\varepsilon=Y-\Delta D$. The result follows by the central limit theorem.

Proof of 2

The proof is very similar as above. For any random variable R, Let $m_{dgt}^{R}(x)=E(R_{dgt}|X=x)$. The estimator satisfies $\widehat{W}_{TC}^{X}=\widehat{N}_{TC}^{X}/\widehat{D}_{TC}^{X}$, with

$$\begin{aligned}
\widehat{N}_{TC}^{X}=&\widehat{E}\left[Y_{11}\right]-\widehat{E}\left[\widehat{m}_{10}^{Y}(X_{11})\right]-\widehat{E}\left[\widehat{m}_{001}^{Y}(X_{11})\right]+\widehat{E}\left[\widehat{m}_{000}^{Y}(X_{11})\right]-\widehat{E}\left[\widehat{m}_{10}^{D}(X_{11})\widehat{m}_{101}^{Y}(X_{11})\right]\\
&+\widehat{E}\left[\widehat{m}_{10}^{D}(X_{11})\widehat{m}_{100}^{Y}(X_{11})\right]+\widehat{E}\left[\widehat{m}_{10}^{D}(X_{11})\widehat{m}_{001}^{Y}(X_{11})\right]-\widehat{E}\left[\widehat{m}_{10}^{D}(X_{11})\widehat{m}_{000}^{Y}(X_{11})\right]\\
\widehat{D}_{TC}^{X}=&\widehat{E}\left[D_{11}\right]-\widehat{E}\left[\widehat{m}_{10}^{D}(X_{11})\right].
\end{aligned}$$

The two terms of the denominator and the first four terms of the numerator can be linearized exactly as above. Regarding the other four terms, remark that for instance

$$\begin{aligned}
&\widehat{E}\left[\widehat{m}_{10}^{D}(X_{11})\widehat{m}_{101}^{Y}(X_{11})\right]-\widehat{E}\left[m_{10}^{D}(X_{11})m_{101}^{Y}(X_{11})\right]\\
=&\widehat{E}\left[m_{10}^{D}(X_{11})\left(\widehat{m}_{101}^{Y}(X_{11})-m_{101}^{Y}(X_{11})\right)\right]+\widehat{E}\left[m_{101}^{Y}(X_{11})\left(\widehat{m}_{10}^{D}(X_{11})-m_{10}^{D}(X_{11})\right)\right]\\
&+\widehat{E}\left[\left(\widehat{m}_{10}^{D}(X_{11})-m_{10}^{D}(X_{11})\right)\left(\widehat{m}_{101}^{Y}(X_{11})-m_{101}^{Y}(X_{11})\right)\right].
\end{aligned}$$

Lemma S7 implies that the last term is an $o_P(1/\sqrt{n})$. As a result,

$$\begin{aligned}
\widehat{N}_{TC}^{X}=&\widehat{E}\left[Y_{11}\right]-\widehat{E}\left[\widehat{m}_{10}^{Y}(X_{11})\right]-\widehat{E}\left[\widehat{m}_{001}^{Y}(X_{11})\right]+\widehat{E}\left[\widehat{m}_{000}^{Y}(X_{11})\right]-\widehat{E}\left[m_{10}^{D}(X_{11})\widehat{m}_{101}^{Y}(X_{11})\right]\\
&-\widehat{E}\left[\widehat{m}_{10}^{D}(X_{11})m_{101}^{Y}(X_{11})\right]+\widehat{E}\left[m_{10}^{D}(X_{11})m_{101}^{Y}(X_{11})\right]+\widehat{E}\left[m_{10}^{D}(X_{11})\widehat{m}_{100}^{Y}(X_{11})\right]\\
&+\widehat{E}\left[\widehat{m}_{10}^{D}(X_{11})m_{100}^{Y}(X_{11})\right]-\widehat{E}\left[m_{10}^{D}(X_{11})m_{100}^{Y}(X_{11})\right]+\widehat{E}\left[m_{10}^{D}(X_{11})\widehat{m}_{001}^{Y}(X_{11})\right]\\
&+\widehat{E}\left[\widehat{m}_{10}^{D}(X_{11})m_{001}^{Y}(X_{11})\right]-\widehat{E}\left[m_{10}^{D}(X_{11})m_{001}^{Y}(X_{11})\right]-\widehat{E}\left[m_{10}^{D}(X_{11})\widehat{m}_{000}^{Y}(X_{11})\right]\\
&-\widehat{E}\left[\widehat{m}_{10}^{D}(X_{11})m_{000}^{Y}(X_{11})\right]+\widehat{E}\left[m_{10}^{D}(X_{11})m_{000}^{Y}(X_{11})\right]+o_P(1/\sqrt{n}).
\end{aligned}$$

We then apply Lemma S8 to each of these terms. After some tedious algebra, we obtain

$$\sqrt{n}\left(\widehat{W}_{TC}^{X}-W_{TC}^{X}\right)=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi_{TC,i}^{X}+o_P(1),$$

where ψ_{TC}^{X} satisfies

$$\begin{aligned}
\psi_{TC}^{X}=&\frac{1}{p_{11}D_{TC}^{X}}\left\{GT\left(U-\Delta(D-m_{10}^{D}(X))-E\left[U_{11}-\Delta(D_{11}-m_{10}^{D}(X_{11}))\right]\right)\right.\\
&\left.+E(GT|X)\left[V-\Delta\frac{G(1-T)}{E(G(1-T)|X)}(D-m_{10}^{D}(X))\right]\right\}. \qquad (47)
\end{aligned}$$

and

$$\begin{aligned}
U=&Y-m_{10}^{Y}(X)-m_{001}^{Y}(X)+m_{000}^{Y}(X)-m_{10}^{D}(X)\left(m_{101}^{Y}(X)-m_{100}^{Y}(X)-m_{001}^{Y}(X)+m_{000}^{Y}(X)\right),\\
V=&\frac{G(1-T)}{E(G(1-T)|X)}\left\{-(Y-m_{10}^{Y}(X))+\left[m_{100}^{Y}(X)-m_{101}^{Y}(X)-m_{000}^{Y}(X)+m_{001}^{Y}(X)\right](D-m_{10}^{D}(X))\right\}\\
&+(1-G)\left\{m_{10}^{D}(X)D\left[\frac{-T(Y-m_{101}^{Y}(X))}{E(D(1-G)T)|X)}+\frac{(1-T)\left(Y-m_{100}^{Y}(X)\right)}{E(D(1-G)(1-T)|X)}\right]\right.\\
&\left.+(1-D)(1-m_{10}^{D}(X))\left[\frac{T(Y-m_{001}^{Y}(X))}{E((1-D)(1-G)T|X)}-\frac{(1-T)(Y-m_{000}^{Y}(X))}{E((1-D)(1-G)(1-T)|X)}\right]\right\}.
\end{aligned}$$

The result follows by the central limit theorem.

Proof of 3

The estimand is the same as W_{TC}^{X}, except for the second term of the numerator. Therefore, it suffices to prove that we can linearize this specific term, which is the plug-in estimator of

$$E\left[E(DQ_{1X}(Y)+(1-D)Q_{0X}(Y)|X,G=1,T=0)|G=1,T=1\right].$$

This expectation comprises two terms. As the reasoning is similar for both, let us focus on the first, $\theta_1 = E\left[E(DQ_{1X}(Y)|X, G=1, T=0)|G=1, T=1\right]$. Let us define $m^{Q_1}_{dgt}(x) = E(Q_{1X}(Y)|X=x, D=d, G=g, T=t)$. First, the estimator $\widehat{\theta}_1$ of θ_1 satisfies

$$\begin{aligned}\widehat{\theta}_1 - \theta_1 &= \widehat{E}\left[\widehat{m}^D_{10}(X)\widehat{m}^{Q_1}_{110}(X)|G=1, T=1\right] - \theta_1 \\ &= \widehat{E}\left[\widehat{m}^D_{10}(X)m^{Q_1}_{110}(X)|G=T=1\right] - \widehat{E}\left[m^D_{10}(X)m^{Q_1}_{110}(X)|G=1, T=1\right] \\ &\quad + \widetilde{\theta}_1 - \theta_1 + \widehat{E}\left[\left(\widehat{m}^D_{10}(X) - m^D_{10}(X)\right)\left(\widehat{m}^{Q_1}_{110}(X) - m^{Q_1}_{110}(X)\right)|G=1, T=1\right], \quad (48)\end{aligned}$$

where $\widetilde{\theta}_1 = \widehat{E}\left[m^D_{10}(X)\widehat{m}^{Q_1}_{110}(X)|G=T=1\right]$. As in parts 1 and 2 above, the first two terms on the right-hand side can be linearized using Lemma S8. We linearize below $\widetilde{\theta}_1 - \theta_1$ and prove that the last term is an $o_P(1/\sqrt{n})$. As in Lemma S5, let us define

$$R_4(F_X, Q_{1|X}, Q_{2|X}, Q_{3|X}) = \int m^D_{10}(x) \times \int_0^1 Q_{1|X}\{Q^{-1}_{2|X}[Q_{3|X}(u|x)|x]|x\}dudF_X(x).$$

Let us define hereafter $F_{dgt|X} = F_{Y_{dgt}|X}$ and $F_{dgt|x} = F_{Y_{dgt}|X=x}$. Because

$$E\left[Q_{1X}(Y)|X=x, D=G=1, T=0\right] = \int_0^1 F^{-1}_{101|x} \circ F_{100|x} \circ F^{-1}_{110|x}(u)du,$$

we have

$$\theta_1 = R_4(F_{X_{11}}, F^{-1}_{101|X}, F^{-1}_{100|X}, F^{-1}_{110|X}), \quad \widetilde{\theta}_1 = R_4(\widehat{F}_{X_{11}}, \widehat{F}^{-1}_{101|X}, \widehat{F}^{-1}_{100|X}, \widehat{F}^{-1}_{110|X}),$$

where $\widehat{F}_{X_{11}}$ is the empirical cdf of X_{11}. By Lemma S9, the process

$$(x, \tau) \mapsto (\widehat{F}_{X_{11}}(x), \widehat{F}^{-1}_{101|x}(\tau), \widehat{F}^{-1}_{100|x}(\tau), \widehat{F}^{-1}_{110|x}(\tau)),$$

defined on $\mathcal{S}(X) \times (0,1)$ and suitably normalized, converges to a continuous gaussian process $\mathbb{G}$. Moreover,

$$\sqrt{n}\left[\widehat{F}^{-1}_{dgt|x}(\tau) - F^{-1}_{dgt|x}(\tau)\right] = \frac{1}{\sqrt{n}}\sum_{i=1}^n \psi_{idgtx}(\tau) + o_P(1),$$

where the $o_P(1)$ is uniform over (x, τ) and

$$\psi_{idgtx}(\tau) = \frac{\mathbb{1}\{D_i = d\}\mathbb{1}\{G_i = g\}\mathbb{1}\{T_i = t\}x'J_\tau X_i}{p_{dgt}}\left(\tau - \mathbb{1}\{Y_i - X_i'\beta_{dgt}(\tau) \leq 0\}\right).$$

Besides, R_4 is Hadamard differentiable at $(F_{X_{11}}, F^{-1}_{101|X}, F^{-1}_{100|X}, F^{-1}_{110|X})$ tangentially to $\mathcal{C}^0(\mathcal{S}(X)) \times \mathcal{C}^0((0,1) \times \mathcal{S}(X))^3$. Therefore, by the functional delta method and because $\mathbb{G}$ is continuous,

$$\sqrt{n}(\widetilde{\theta}_1 - \theta_1) = \frac{1}{\sqrt{n}}\sum_{i=1}^n \Psi_{1i} + o_P(1),$$

where

$$\Psi_{1i} = \frac{G_i T_i}{p_{11}}\left[m_{10}^D(X_i)m_{110}^{Q_1}(X_i) - \theta_1\right] + \int m_{10}^D(x)\left\{\int_0^1 \psi_{i101x}\left(F_{100|x}\circ F_{110|x}^{-1}(u)\right)\right.$$
$$\left. + \frac{{F_{101|x}^{-1}}'\circ F_{100|x}\circ F_{110|x}^{-1}(u)}{{F_{100|x}^{-1}}'\circ F_{100|x}\circ F_{110|x}^{-1}(u)}\left[-\psi_{i100x}\left(F_{100|x}\circ F_{110|x}^{-1}(u)\right) + \psi_{i110x}(u)\right]du\right\} dF_{X_{11}}(x).$$

We now prove that the third term in (48) is an $o_P(1/\sqrt{n})$. We have

$$\left|\widehat{E}\left[\left(\widehat{m}_{10}^D(X) - m_{10}^D(X)\right)\left(\widehat{m}_{110}^{Q_1}(X) - m_{110}^{Q_1}(X)\right)|G=1,T=1\right]\right|$$
$$\leq \left\|\widehat{m}_{10}^D - m_{10}^D\right\|_\infty \times \left\|\widehat{m}_{110}^{Q_1} - m_{110}^{Q_1}\right\|_\infty.$$

By Lemma S7, $\left\|\widehat{m}_{10}^D - m_{10}^D\right\|_\infty = o_P(n^{-1/4})$. Besides, $\widehat{m}_{110}^{Q_1} = R_5(\widehat{F}_{101|X}^{-1}, \widehat{F}_{100|X}^{-1}, \widehat{F}_{110|X}^{-1})$, where $R_5(Q_{1|X}, Q_{2|X}, Q_{3|X}) = \int_0^1 Q_{1|X}\{Q_{2|X}^{-1}[Q_{3|X}(u|x)|x]|x\}du$. Part 3 of the proof of Lemma S5 implies that R_5 is Hadamard differentiable at $(F_{101|X}^{-1}, F_{100|X}^{-1}, F_{110|X}^{-1})$. Then, by Lemma S9 and the functional delta method, $\left\|\widehat{m}_{110}^{Q_1} - m_{110}^{Q_1}\right\|_\infty = O_P(n^{-1/2})$. Thus, the third term in (48) is an $o_P(1/\sqrt{n})$.

To conclude, we provide the linearization of W_{CIC}^X. Let us define for that purpose

$$\Psi_{0i} = \frac{G_i T_i}{p_{11}}\left[(1 - m_{10}^D(X_i))m_{010}^{Q_0}(X_i) - \theta_0\right] + \int (1 - m_{10}^D(x))\left\{\int_0^1 \psi_{i001x}\left(F_{000|x}\circ F_{010|x}^{-1}(u)\right)\right.$$
$$\left. + \frac{{F_{001|x}^{-1}}'\circ F_{000|x}\circ F_{010|x}^{-1}(u)}{{F_{000|x}^{-1}}'\circ F_{000|x}\circ F_{010|x}^{-1}(u)}\left[-\psi_{i000x}\left(F_{000|x}\circ F_{010|x}^{-1}(u)\right) + \psi_{i010x}(u)\right]du\right\} dF_{X_{11}}(x),$$

where $\theta_0 = E\left[E((1-D)Q_{0X}(Y)|X, G=1, T=0)|G=1,T=1\right]$. Using what precedes and Lemma S8 on the remaining terms, we obtain after some tedious algebra

$$\sqrt{n}\left(\widehat{W}_{CIC}^X - W_{CIC}^X\right) = \frac{1}{\sqrt{n}}\sum_{i=1}^n \psi_{CIC,i}^X + o_P(1),$$

where ψ_{CIC}^X satisfies

$$\psi_{CIC}^X = \frac{1}{p_{11}D_{CIC}^X}\left\{GT\left(Y - \Delta(D - m_{10}^D(X)) - E\left[Y_{11} - \Delta(D_{11} - m_{10}^D(X_{11}))\right]\right) - p_{11}(\Psi_1 + \Psi_0)\right.$$
$$\left. + \frac{E(GT|X)G(1-T)}{E(G(1-T)|X)}(D - m_{10}^D(X))\left[m_{010}^{Q_0}(X) - m_{110}^{Q_1}(X) - \Delta\right]\right\}. \quad (49)$$