Fuzzy differences in differences

Clément de Chaisemartin

PARIS SCHOOL OF ECONOMICS
ÉCOLE D'ÉCONOMIE DE PARIS

WORKING PAPER N° 2012 – 06

Fuzzy differences in differences

Clément de Chaisemartin

JEL Codes: C21, C23, I19

Keywords: Difference in Differences, Heterogeneous Treatment Effect, Imperfect Compliance, Partial Identification, Smoking Cessation

PARIS-JOURDAN SCIENCES ECONOMIQUES
48, BD JOURDAN – E.N.S. – 75014 PARIS
TÉL. : 33(0) 1 43 13 63 00 – FAX : 33 (0) 1 43 13 63 10
www.pse.ens.fr

CENTRE NATIONAL DE LA RECHERCHE SCIENTIFIQUE – ECOLE DES HAUTES ETUDES EN SCIENCES SOCIALES
ÉCOLE DES PONTS PARISTECH – ECOLE NORMALE SUPÉRIEURE – INSTITUT NATIONAL DE LA RECHERCHE AGRONOMIQUE

Fuzzy Differences in Differences *

Clément de Chaisemartin†

February 15, 2012

Abstract

Difference in differences require that 0% of observations are treated in the control group and during period 0 (no "always takers") and 100% in the treatment group in period 1 (no "never takers"). Sometimes, the treatment rate increases more in the treatment than in the control group but there are never or always takers. This paper develops results to identify treatment effects in such settings. They only require one common trend assumption on the outcome of interest Y whereas the standard instrumental variable result also requires common trend on treatment D. I derive bounds for treatment effects which are tight when there are no or few always takers. This can be the case in applications considering the effect of an innovation, where by definition no observations are treated in period 0. I derive other bounds that are tight when the treatment rate does not change much between the two periods in the control group, which can be the case in applications considering the extension of a program to a group previously not eligible. I use my results to measure the efficacy of a new drug for smoking cessation.

Keywords: Difference in differences, heterogeneous treatment effect, imperfect compliance, partial identification, smoking cessation

JEL Codes: C21, C23, I19

*I am very grateful to Sandra Black, Christian Bontemps, Joseph Doyle, Pauline Givord, Marc Gurgand, Xavier d'Haultfoeuille, Edwin Leuven, Thierry Magnac, Eric Maurin, Thomas Piketty, Roland Rathelot, Eric Verhoogen, participants at the French Econometrics Conference, the European Conference in Econometrics (EC^2), the 7th IZA Conference on Labor Market Policy Evaluation, the North American and European Summer Meetings of the Econometric Society, the 11th Doctoral Workshop in Economic Theory and Econometrics, seminar participants at the Paris School of Economics and at CREST for their helpful comments. I gratefully acknowledge financial support from "Région Ile de France" for this research.

†Paris School of Economics (48 boulevard Jourdan 75014 Paris France) and CREST (15 boulevard Gabriel Péri 92245 Malakoff France), chaisemartin@pse.ens.fr

1 Introduction

Differences in differences (DID) are commonly used to estimate average treatment effects on the treated (ATT) when treatment D is not randomly allocated. DID compare the evolution of the mean of some outcome Y between two periods (0 and 1) and across two groups of individuals (control and treatment). In Rubin's causal model (Rubin (1974)) where potential outcomes with and without treatment ($Y(1)$ and $Y(0)$) are introduced, and where treatment effects are allowed to be heterogeneous across observations, a DID identifies an ATT under two assumptions. The first one is a common trend assumption which states that if all observations had remained untreated the mean of Y would have followed parallel trends from period 0 to 1 in the two groups (see e.g. Abadie, 2005). The second one, which is implicit, is a perfect compliance assumption: the treatment rate should be equal to 0% in the control group and during period 0 (no "always takers") and to 100% in the treatment group in period 1 (no "never takers").[1] In many instances, this last assumption is violated: the treatment rate increases more in the treatment than in the control group but there are "never" or "always" takers.[2] This differential change in treatment rate across the control and the treatment group might still be used to identify an ATT.

The starting point of the paper is that when compliance is imperfect, common trend alone is not sufficient for identification in a model allowing for heterogeneous treatment effects. In a standard DID, if the common trend assumption on $Y(0)$ holds, the only reason why trends might diverge across groups is that observations in the treatment group $\times$ period 1 cell get treated, so that the DID measures the effect of the treatment on them. A DID computation will therefore yield one equation with only one unknown. In a fuzzy DID, since there might be treated observations in each of the four time $\times$ group cells, diverging trends can potentially arise from the effect of the treatment within each of those four subgroups and a DID computation will yield one equation with up to four unknowns. The identification problem arises because $Y(1) - Y(0)$ is allowed to vary across observations, implying that the

[1] In this paper, never takers are merely untreated observations in the period 1 $\times$ treatment group cell. Always takers are treated observations in the three other cells. Therefore, my definition of always and never takers does not exactly correspond to the definition of Imbens & Angrist (1994), but I still use their terminology to clarify the exposition.

[2] When panel data is available, one option to avoid this issue could be to use treatment status in period 0 and 1 to define the treatment and the control groups, instead of using presumably stable observable characteristics. One could for instance define all observations untreated in period 0 and 1 to be the control group, and all observations untreated in period 0 and treated in period 1 to be the treatment group (see for instance Field (2005)). However, when groups are constructed this way, common trend might not hold, for instance because untreated individuals in period 0 may decide to get treated in period 1 if they expect a negative shock on their $Y(0)$.

effect of the treatment might vary across cells.

However, when there are no always takers, I show in Theorem 2.1 that the common trend assumption is sufficient to identify an ATT because there are treated observations in one group only, so that the DID computation yields one equation with one unknown. Applications with no always takers arise frequently in practice. They correspond to situations where it is possible to isolate one group excluded from treatment both in period 0 and 1 based on observable characteristics, while it is not possible to isolate a group fully treated in period 1. This can happen for various reasons: eligibility criteria might be too complex to determine exactly the eligible population, or some eligible individuals might deny the treatment. A good example is Eissa & Liebman (1996).[3]

When there are always takers, common trend on $Y(0)$ does not allow for point identification, but partial identification of an ATT in the spirit of Manski (1990) is still possible, provided Y is bounded. In Theorem 2.2, I derive sharp bounds in this case. Those bounds will be tight when there are "few" always takers. I illustrate what "few" means in the next section through a simple numerical example. For now, one can keep in mind as a rough rule of thumb that those bounds can be relatively tight when the sum of the shares of observations treated in the three supposedly untreated cells is below 10%. This is likely to happen in applications considering the effect of an innovation released in period 1. Indeed, in such applications nobody is treated in period 0. Therefore, bounds will be tight if it is possible to find a control group in which less than 10% of observations benefited from the innovation in period 1. A good example is the application used in this paper. Varenicline is a drug which was made available to French smoking cessation clinics in February 2007. In 15 clinics, less than 3% of all patients consulted have been prescribed varenicline during the year following its release. In 13 clinics, more than 20% of patients received it. Overall, 38% of patients received this new drug in the treatment clinics in period 1, against 2% in the control clinics. I use Theorem 2.2 to derive bounds for the average effect of varenicline on smoking cessation. Since in this application there are very few always takers, those bounds are narrow.

Then, I derive a second couple of sharp bounds for the same ATT under the supplementary assumption that the response to treatment is monotone (i.e. $Y(1) \geq Y(0)$) as in Manski (1997).

[3]To measure the effect of EITC extension on female labor market participation, Eissa & Liebman (1996) use single women without children as a control group, and lone mothers as the treatment group. By definition, their control group is excluded from treatment both in period 0 and 1. Moreover, treatment is not available to the treatment group in period 0. Therefore, they do not have always takers in their analysis. However, not all lone mothers are eligible to EITC extension in period 1 because this program is means tested. Exact eligibility criteria are too complex for them to isolate eligible lone mothers. Moreover, it is likely that not all lone mothers eligible applied for EITC. Therefore, they have never takers.

Those bounds partly dispose of the bounded support assumption since they only require that $Y(0)$ is bounded by below. Therefore, they apply for instance to wages, while the previous ones do not. The size of the resulting identification region also depends on the shares of always takers. I show in a numerical example that it is roughly twice smaller than the first one.

Finally, I derive a third couple of sharp bounds under the assumption that treatment effects do not change between the two periods in the control group. The size of the resulting identification region now depends on the share of treated observations in the treatment group in period 0, and on the change in the treatment rate from period 0 to 1 in the control group. I show through a numerical example that when there are no treated observations in the treatment group in period 0, those bounds will be relatively tight as long as the treatment rate does not change by more than roughly 10 percentage points in the control group. This is likely to be the case in applications considering the extension of a public policy to a group previously not eligible to it, and where the previously eligible group is used as a control group (see for instance Bach, 2010).

Common trend is a strong identifying assumption which might appear implausible if pre-treatment characteristics that are thought to be associated to the dynamics of the outcome variable are unbalanced between the treatment and the control groups (see Abadie, 2005). In such instances, a conditional common trend assumption might appear more credible. Therefore, I also derive fuzzy semi-parametric DID results inspired from Abadie (2005), which allow to control semi-parametrically for covariates. All the results obtained under common trend can be extended under conditional common trend.

Many empirical papers have already used a greater increase of the treatment rate in one group as a source of variation to identify treatment effects. Up to now, researchers who implemented this strategy estimated the impact of the treatment through an instrumental variable (IV) regression using the interaction of time and group as an instrument for treatment. The resulting coefficient is the DID on Y divided by the DID on D. I hereafter refer to this strategy as an IV-DID. Duflo (2001) uses an IV-DID to estimate the impact of educational attainment on wages. There are also a very large number of papers which use differential evolution of exposure to treatment across US states to estimate treatment effects. A good example is Evans & Ringel (1999) who use changes in cigarette taxes across US states as an instrument for smoking prevalence among pregnant women, in order to estimate the impact of smoking during pregnancy on newborns' weight.

Imbens & Angrist (1994) have shown that IV coefficients can be interpreted as local average treatment effects (LATE) in a model allowing for heterogeneous treatment effects. I put forward in a companion note (de Chaisemartin (2011)) that when applied to IV-DID, their

result holds under a common trend assumption on $Y(0)$ but also on $D(0)$ (the potential treatment without the instrument), and a monotonicity assumption which states that there should be no "defiers". Common trend on $Y(0)$ ensures that the DID on Y recovers the intention to treat effect of the policy, whereas common trend on $D(0)$ and monotonicity ensure that the DID on D is equal to the share of compliers, so that the ratio of the two is indeed equal to a LATE.

My results contribute to the literature firstly because they remove the monotonicity condition which may be restrictive in some applications.[4] More importantly, my results hold under a common trend assumption on $Y(0)$ only, and not on $D(0)$. There are two reasons why disposing of the common trend assumption on $D(0)$ while maintaining it on $Y(0)$ might be appealing. Firstly, in many IV-DID applications, common trend on $D(0)$ is less credible than common trend on $Y(0)$. For instance, the identifying assumption in Evans & Ringel (1999) is that both mothers' smoking rates and newborns' weight would have followed parallel trends in states where taxes on tobacco increased and in other states if taxes had not increased. But it may be the case that states which choose to rise taxes on cigarettes do so because they face an increasing trend in smoking, whereas there is no reason to suspect that this decision is related to trends on other determinants of newborns weight.

Secondly, there are applications in which taking a common trend assumption on $D(0)$ is merely problematic. For instance, in applications considering the extension of a public policy, common trend on $D(0)$ means that if the policy had not been extended to the treatment group in period 1, the treatment rate would have followed the same trends in the control and in the treatment groups. But if the policy had not been extended, the treatment rate would merely have been equal to 0% at both periods in the treatment group. Therefore, common trend on $D(0)$ will be violated as soon as the treatment rate slightly changes in the control group. In applications considering the introduction of an innovation, common trend on $D(0)$ is problematic as well. Take for instance the application developed in this paper. Treatment rate was equal to 0% in the two groups of clinics in period 0. Therefore, common trend on $D(0)$ means that if treatment group patients had gone to a control clinic in period 1, 2% of them would have received varenicline as well, exactly as what indeed happened to control group patients. Since allocation of patients to clinics was not random, this assumption is unlikely to hold.

[4] A good example is Angrist & Evans (1998), where the authors use the fact that parents show preferences for a mixed sibling-sex composition as an instrument to estimate the effect of childbearing on females labor supply. Even though parents might on average display such preferences, it is not impossible that some parents have the opposite preferences, so that the monotonicity assumption might hold on average but not with probability one.

Importantly, the bounds identified in this paper are very simple to estimate. They can be estimated through OLS regressions on slightly modified versions of Y, which essentially amount to replace always takers' Y by the lower or upper endpoint of the support of $Y(0)$. Inference is more involved, because of all the complications arising when conducting inference on partially identified parameters (see e.g. Imbens & Manski (2004)). However, a recent method developed by Andrews & Soares (2010) readily applies to my setting.

The remainder of the paper is organized as follows. Section 2 is devoted to identification. Section 3 deals with inference. Section 4 is devoted to the application. Section 5 concludes.

2 Identification

Let $T \in \{t_0; t_1\}$ denote time and $G \in \{g_c; g_t\}$ denote treatment (g_t) and control (g_c) groups. Results apply irrespective of whether panel or pooled cross-sections data are available. Treatment status is binary and is denoted by an indicator D.[5] Let $Y(1)$ and $Y(0)$ be the potential outcomes of an individual with and without the treatment. Only the actual outcome $Y = Y(1) \times D + Y(0) \times (1 - D)$ is observed. The individual treatment effect is $Y(1) - Y(0)$.

For any random variables Z and W, $Z \sim W$ means that Z and W have the same probability distribution. $\mathbb{Z}$ is the support of Z. To alleviate the notational burden, I introduce the following shorthand taken from Athey & Imbens (2006):

$$\forall (i,j) \in \{t_0; t_1\} \times \{g_c; g_t\}, \quad Z_{i,j} \sim Z |\, t = i, g = j.$$

Under those notations, $\forall (i,j) \in \{t_0; t_1\} \times \{g_c; g_t\}$, $ATT_{i,j} = \mathbb{E}(Y_{i,j}(1) - Y_{i,j}(0) |\, D = 1)$ is the average treatment effect on treated individuals of group j in period i. My parameter of interest is ATT_{t_1,g_t}, which is the average treatment effect among treated observations in the $(t_1; g_t)$ cell. This population is directly identified from the data (provided D is observed) and can be easily characterized.

The DID of a random variable Z is denoted

$$DID_Z = \mathbb{E}(Z_{t_1,g_t}) - \mathbb{E}(Z_{t_0,g_t}) - [\mathbb{E}(Z_{t_1,g_c}) - \mathbb{E}(Z_{t_0,g_c})].$$

I assume that DID_D, the DID on treatment rate, is different from 0: treatment rate should not follow the same evolution in the two groups. Without loss of generality, I assume that $DID_D > 0$. I denote

$$\mathbb{P}_{AT} = \mathbb{P}(D_{t_0,g_t} = 1) + \mathbb{P}(D_{t_1,g_c} = 1) + \mathbb{P}(D_{t_0,g_c} = 1)$$

[5] The analysis can easily be extended when treatment is multivariate in the spirit of Angrist & Imbens (1995). Results are not presented here due to a concern for brevity but are available upon request.

the sum of the three shares of always takers. The "no always takers" special case is when $\mathbb{P}_{AT} = 0$, i.e. when there are no treated observations in any of the supposedly untreated cells. As detailed in the introduction, this special case is important in practice.

I take a common trend assumption which is at the basis of the DID approach (see for instance Abadie (2005)):

A.1 Common trend

$$\mathbb{E}(Y_{t_1,g_t}(0)) - \mathbb{E}(Y_{t_0,g_t}(0)) = \mathbb{E}(Y_{t_1,g_c}(0)) - \mathbb{E}(Y_{t_0,g_c}(0)).$$

This common trend assumption can be rationalized by a DGP for potential outcomes additively separable in time and group:

$$Y(d) = f_d(T, U) + g_d(G, U),$$

where U represents unobserved heterogeneity and $U \perp\!\!\!\perp (T, G)$ (see e.g. Athey & Imbens (2006)). I can now state the lemma which is at the core of all results in the paper:

Lemma 2.1 ***Non-identification***

Under A.1, none of the $ATT_{i,j}$ is identified and

$$\begin{aligned} DID_Y = ATT_{t_1,g_t} \times \mathbb{P}(D_{t_1,g_t} = 1) - ATT_{t_0,g_t} \times \mathbb{P}(D_{t_0,g_t} = 1) \\ -ATT_{t_1,g_c} \times \mathbb{P}(D_{t_1,g_c} = 1) + ATT_{t_0,g_c} \times \mathbb{P}(D_{t_0,g_c} = 1). \end{aligned} \tag{1}$$

According to Lemma 2.1, under A.1 DID_Y can be written as a weighted DID of four average treatment effects. Because two ATT enter the equation with positive sign and two with negative sign, DID_Y cannot be given any causal interpretation. It might for instance be positive whereas the four ATT are negative.

The intuition for this result is as follows. According to A.1, if no observations had been treated in any of the four time $\times$ group cells, trends would have been parallel in the two groups, and DID_Y would have merely been equal to 0. In a standard DID, if A.1 holds then the only reason why trends might diverge across groups is that observations in the treatment group get treated in period 1, so that DID_Y measures the effect of the treatment on them. In a fuzzy DID, since there might be treated observations in several time $\times$ group cells, diverging trends can potentially arise from the effect of the treatment in each of those cells. Then, if no restrictions are placed on how heterogeneous the treatment effect can be across these four cells, it is not possible to identify any of the $ATT_{i,j}$ from a standard DID computation, since it yields one equation with four unknowns. From this Lemma, I derive Theorems 2.1 to 2.4.

Theorem 2.1 ***Point identification***

1. *Under A.1, in the no always takers special case,*

$$ATT_{t_1,g_t} = \frac{DID_Y}{\mathbb{P}(D_{t_1,g_t} = 1)}.$$

2. *Under A.1, if* $\forall (i,j) \in \{t_0; t_1\} \times \{g_c; g_t\}$, $ATT_{i,j} = ATT_{t_1,g_t}$, ATT_{t_1,g_t} *is identified:*

$$ATT_{t_1,g_t} = \frac{DID_Y}{DID_D}.$$

In the no always takers special case, there are treated observations in one group only, and only one unknown left in (1). Therefore, A.1 is sufficient to identify an average treatment effect, as in a standard DID.[6] This average treatment effect is ATT_{t_1,g_t}, the average effect of the treatment among treated observations in the $(t_1; g_t)$ cell. Estimating it still requires being able to estimate $\mathbb{P}(D_{t_1,g_t} = 1)$. But sometimes treatment status is not observed. In such instances, since ATT_{t_1,g_t} and DID_Y have the same sign and $|DID_Y| \leq |ATT_{t_1,g_t}|$, it is at least possible to estimate a lower bound of ATT_{t_1,g_t} by computing DID_Y. For instance, Eissa & Liebman (1996) do not observe treatment status, so that they cannot estimate $\frac{DID_Y}{\mathbb{P}(D_{t_1,g_t}=1)}$. They estimate instead DID_Y, and find that participation to the labor market increased by 1.4 percentage points more among lone mothers than among single women without children following the extension of the EITC. The first point in Theorem 2.1 allows to interpret this 1.4 percentage points DID as a lower bound to the true effect of the EITC extension on labor market participation among lone mothers who benefited from it.

The second point of Theorem 2.1 shows that even when there are always takers, ATT_{t_1,g_t} is also identified if one is ready to assume that treatment effects do not vary across time and group. This is because under this assumption the four unknowns in (1) are equal to each other. But this is fairly restrictive an assumption. The underlying assumption to a fuzzy DID is indeed that treatment rate increased more from period 0 to 1 in the treatment group than in the control group. This might for instance be because treatment group individuals were more incentivized to receive the treatment in period 1 than in period 0. Inside the treatment group, treated individuals during period 1 are therefore likely to differ from those treated during period 0 so that the average treatment effect could arguably be different in these two

[6] This is similar to the result on regression discontinuity (RDD) obtained by Battistin & Rettore (2008). They indeed show that in a fuzzy RDD, when treatment rate is equal to 0 below the eligibility threshold, so that fuzziness arises only because of never takers (i.e. untreated individuals above the threshold), identification is obtained under the same assumptions than in a sharp RDD.

groups. Therefore, I give now three partial identification results which do not require this "constant treatment effect" assumption.

After some algebra, (1) rewrites as

$$ATT_{t_1,g_t} = \frac{\mathbb{E}(Y_{t_1,g_t}) - \mathbb{E}(Y_{t_0,g_t}(0)) - \mathbb{E}(Y_{t_1,g_c}(0)) + \mathbb{E}(Y_{t_0,g_c}(0))}{\mathbb{P}(D_{t_1,g_t} = 1)}. \tag{2}$$

This implies that ATT_{t_1,g_t} would be identified if the mean of $Y(0)$ was observed in the $(t_0; g_t)$, $(t_1; g_c)$ and $(t_0; g_c)$ cells. Therefore ATT_{t_1,g_t} is not identified when there are treated observations in those cells, because their $Y(0)$ is not observed. But when $Y(0)$ is bounded, it is possible to bound those unobserved $Y(0)$ to derive some bounds for ATT_{t_1,g_t}. This is what I do in Theorem 2.2.

Before stating Theorem 2.2, I define three indicator variables. $AT_{t_0,g_t} = 1_{\{D=1,T=t_0,G=g_t\}}$ corresponds to treated observations in the $(t_0; g_t)$ cell, $AT_{t_1,g_c} = 1_{\{D=1,T=t_1,G=g_c\}}$ to treated observations in the $(t_1; g_c)$ cell, and $AT_{t_0,g_c} = 1_{\{D=1,T=t_0,G=g_c\}}$ to treated observations in the $(t_0; g_c)$ cell. I define the following random variables which are functions of Y, D, T, G, M and m, where m and M are real numbers defined below:

$$Y^0_- = Y + (M - Y)\left(AT_{t_0,g_t} + AT_{t_1,g_c}\right) + (m - Y)AT_{t_0,g_c}$$

and

$$Y^0_+ = Y + (m - Y)\left(AT_{t_0,g_t} + AT_{t_1,g_c}\right) + (M - Y)AT_{t_0,g_c}.$$

Y^0_- merely replaces Y by M for treated observations in the $(t_0; g_t)$ and $(t_1; g_c)$ cells, and by m for treated observations in the $(t_0; g_c)$ cell, while it leaves Y unchanged for all remaining observations. Y^0_+ performs the same replacements except that M and m are switched.

Theorem 2.2 ***Partial identification 1***

Under A.1, if $\mathbb{P}(m \leq Y(0) \leq M) = 1$, *with* $(m, M) \in \mathbb{R}^2$,

$$B_- \leq ATT_{t_1,g_t} \leq B_+,$$

with

$$B_- = \max\left(\mathbb{E}(Y_{t_1,g_t} | \, D = 1) - M; \frac{DID_{Y^0_-}}{\mathbb{P}(D_{t_1,g_t} = 1)}\right)$$

and

$$B_+ = \min\left(\mathbb{E}(Y_{t_1,g_t} | \, D = 1) - m; \frac{DID_{Y^0_+}}{\mathbb{P}(D_{t_1,g_t} = 1)}\right).$$

B_- *and* B_+ *are sharp.*

B_- is obtained as follows. Firstly, boundedness of $Y(0)$ trivially implies that

$$\mathbb{E}(Y_{t_1,g_t}|\, D=1) - M \leq ATT_{t_1,g_t}.$$

Secondly, the only unobserved quantities in (2) are $\mathbb{E}(Y_{t_0,g_t}(0)|\, D=1)$, $\mathbb{E}(Y_{t_1,g_c}(0)|\, D=1)$ and $\mathbb{E}(Y_{t_0,g_c}(0)|\, D=1)$. Therefore, using the fact that $m \leq \mathbb{E}(Y_{t_0,g_t}(0)|\, D=1)$, $m \leq \mathbb{E}(Y_{t_1,g_c}(0)|\, D=1)$ and $\mathbb{E}(Y_{t_0,g_c}(0)|\, D=1) \leq M$, and plugging those three inequalities into (2) yields another lower bound for ATT_{t_1,g_t}. Finally, B_- is defined as the max of those two lower bounds.

Interestingly, one can show that the second lower bound writes as

$$\frac{DID_{Y_-^0}}{\mathbb{P}(D_{t_1,g_t}=1)}.$$

This expression is similar to the formula obtained in point 1 of Theorem 2.1, except that the DID is no longer computed on Y but on Y_-^0. Y_-^0 replaces Y by M for treated observations in the $(t_0; g_t)$ and $(t_1; g_c)$ cells, and by m for treated observations in the $(t_0; g_c)$ cell. Therefore, $DID_{Y_-^0}$ is the lowest possible value of the numerator of equation (2). $DID_{Y_-^0}$ can be estimated through the following OLS regression:

$$Y_-^0 = \alpha + \beta T + \gamma G + \theta TG + u.$$

Therefore, estimation of B_- is fairly straightforward.

The first lower and upper bounds, i.e. $\mathbb{E}(Y_{t_1,g_t}|\, D=1) - M$ and $\mathbb{E}(Y_{t_1,g_t}|\, D=1) - m$, arise from boundedness of $Y(0)$ alone, while

$$\frac{DID_{Y_-^0}}{\mathbb{P}(D_{t_1,g_t}=1)}$$

and

$$\frac{DID_{Y_+^0}}{\mathbb{P}(D_{t_1,g_t}=1)}$$

arise from the combination of boundedness and common trend. When $B_- = \mathbb{E}(Y_{t_1,g_t}|\, D=1) - M$ and $B_+ = \mathbb{E}(Y_{t_1,g_t}|\, D=1) - m$, the bounds are uninformative: common trend does not add any supplementary information on ATT_{t_1,g_t} to the information we can derive from the mere fact that $Y(0)$ is bounded. It is easy to show that if $\mathbb{P}_{AT} \leq \mathbb{P}(D_{t_1,g_t}=1)$, that is to say if the share of treated observations in the $(t_1; g_t)$ cell is greater than the sum of the shares of always takers, then at least one of the bounds is informative. Conversely, when $\mathbb{P}_{AT} > \mathbb{P}(D_{t_1,g_t}=1)$, at least one of the bounds is uninformative.

When

$$B_- = \frac{DID_{Y_-^0}}{\mathbb{P}(D_{t_1,g_t}=1)}$$

and

$$B_+ = \frac{DID_{Y_+^0}}{\mathbb{P}(D_{t_1,g_t}=1)},$$

the length of $[B_-; B_+]$ is equal to

$$(M-m) \times \frac{\mathbb{P}_{AT}}{\mathbb{P}(D_{t_1,g_t}=1)}.$$

Therefore, the two bounds will be tight when $\frac{\mathbb{P}_{AT}}{\mathbb{P}(D_{t_1,g_t}=1)}$ is small, that is to say when there are "few" always takers with respect to treated observations in the $(t_1; g_t)$ cell.

I illustrate what "few" always takers means through a numerical example. I consider a binary outcome Y, such that $DID_Y = 0.05$: the mean of Y increased by 5 percentage points more in the treatment than in the control group from period 0 to 1. To simplify the discussion, I assume that $\mathbb{P}(D_{t_0,g_t}=1) = \mathbb{P}(D_{t_0,g_c}=1) = 0$, so that $\mathbb{P}_{AT} = \mathbb{P}(D_{t_1,g_c}=1)$. This corresponds for instance to applications considering an innovation released in period 1 so that no observation is treated in period 0 in the treatment and in the control groups. Finally, I assume that $\mathbb{P}(D_{t_1,g_t}=1) = 0.5$ and $\mathbb{E}(Y_{t_1,g_c}|\, D=1) = 0.6$. Under those assumptions, I can write B_- and B_+ as functions of $\mathbb{P}(D_{t_1,g_c}=1)$, the share of treated observations in the control group in period 1. Results are summarized in Table 1.

Table 1: Value of B_- and B_+ according to $\mathbb{P}(D_{t_1,g_c}=1)$

$\mathbb{P}(D_{t_1,g_c}=1)$	0%	2.5%	5%	7.5%	10%	12.5%	15%	17.5%	20%
B_-	10%	8%	6%	4%	2%	0%	-2%	-4%	-6%
B_+	10%	13%	16%	19%	22%	25%	28%	31%	34%

The length of $[B_-; B_+]$ linearly increases with $\mathbb{P}(D_{t_1,g_c}=1)$, from 0% at $\mathbb{P}(D_{t_1,g_c}=1)=0$, to 40% at $\mathbb{P}(D_{t_1,g_c}=1)=0.2$. The sign of ATT_{t_1,g_t} is identified as long as $\mathbb{P}(D_{t_1,g_c}=1)$ is below 12.5%.

Then, I show in Theorem 2.3 that it is possible to derive tighter bounds under the monotone treatment response assumption considered by Manski (1997). Before stating Theorem 2.3, I define the following random variables:

$$Y_-^1 = Y + (m-Y)AT_{t_0,g_c}$$

and

$$Y_+^1 = Y + (m-Y)\left(AT_{t_0,g_t} + AT_{t_1,g_c}\right).$$

Y_-^1 replaces Y by m for treated observations in the $(t_0; g_c)$ cell, while Y_+^1 replaces Y by m for treated observations in the $(t_0; g_t)$ and $(t_1; g_c)$ cells.

Theorem 2.3 ***Partial identification 2***

Under A.1, if $\mathbb{P}(m \leq Y(0) \leq Y(1)) = 1$ *with* $m \in \mathbb{R}$,

$$B_{-}^{'} \leq ATT_{t_1,g_t} \leq B_{+}^{'},$$

with

$$B_{-}^{'} = \max\left(0; \frac{DID_{Y_{-}^{1}}}{\mathbb{P}(D_{t_1,g_t}=1)}\right)$$

and

$$B_{+}^{'} = \min\left(\mathbb{E}(Y_{t_1,g_t}|\, D=1) - m; \frac{DID_{Y_{+}^{1}}}{\mathbb{P}(D_{t_1,g_t}=1)}\right).$$

$B_{-}^{'}$ *and* $B_{+}^{'}$ *are sharp.*

The monotone treatment response assumption under which I derive 2.3 states that the effect of the treatment on the outcome is always positive.[7] As emphasized in Manski (1997), it might be credible in some economic contexts, such as the analysis of the supply of a good which can be assumed to be an increasing function of its price. This assumption partly disposes of the bounded support assumption which was requested in Theorem 2.2, since in Theorem 2.3 it is only requested that the support of $Y(0)$ is bounded by below. Therefore, Theorem 2.3 applies for instance when Y are wages while Theorem 2.2 does not. $B_{-}^{'}$ is obtained through the exact same steps as B_{-}, except that the monotonicity assumption allows me to use $Y(1)$ instead of M as an upper bound for always takers' $Y(0)$. Consequently, $\left[B_{-}^{'}; B_{+}^{'}\right]$ is shorter than $[B_{-}; B_{+}]$. I compute $B_{-}^{'}$ and $B_{+}^{'}$ in the same numerical example as above to illustrate to what extent the monotonicity assumption tightens the bounds. One can see in Table 2: $\left[B_{-}^{'}; B_{+}^{'}\right]$ is approximately twice smaller than $[B_{-}; B_{+}]$.

Table 2: Value of $B_{-}^{'}$ and $B_{+}^{'}$ according to $\mathbb{P}(D_{t_1,g_c}=1)$

$\mathbb{P}(D_{t_1,g_c}=1)$	0%	2.5%	5%	7.5%	10%	12.5%	15%	17.5%	20%
$B_{-}^{'}$	10%	10%	10%	10%	10%	10%	10%	10%	10%
$B_{+}^{'}$	10%	13%	16%	19%	22%	25%	28%	31%	34%

Finally, it is possible to derive bounds tighter than those obtained in Theorem 2.2 under the assumption that the effect of the treatment does not change between period 0 and 1 in the control group.[8] Before stating Theorem 2.4, I introduce new notations.

[7] One could also derive another partial identification result under the alternative assumption that the effect of the treatment is always negative.

[8] I am very grateful to Roland Rathelot for suggesting this result.

Let $1_{\{t_1\}}$ (resp. $1_{\{t_0\}}$) be an indicator equal to 1 when $\mathbb{P}(D_{t_0,g_c}=1) < \mathbb{P}(D_{t_1,g_c}=1)$ (resp. $\mathbb{P}(D_{t_0,g_c}=1) > \mathbb{P}(D_{t_1,g_c}=1)$). Let $\Delta\mathbb{E} = \mathbb{E}(Y_{t_1,g_c}|D=1) - \mathbb{E}(Y_{t_0,g_c}|D=1)$. Let

$$\begin{aligned}
Y_-^2 &= Y + AT_{t_0,g_t}(M-Y) \\
&+AT_{t_1,g_c}\left(1_{\{t_1\}}\left(\min\left(M; M+\Delta\mathbb{E}\right)-Y\right) + 1_{\{t_0\}}\left(\max\left(m; m+\Delta\mathbb{E}\right)-Y\right)\right) \\
&+AT_{t_0,g_c}\left(1_{\{t_1\}}\left(\min\left(M; M-\Delta\mathbb{E}\right)-Y\right) + 1_{\{t_0\}}\left(\max\left(m; m-\Delta\mathbb{E}\right)-Y\right)\right)
\end{aligned}$$

and

$$\begin{aligned}
Y_+^2 &= Y + AT_{t_0,g_t}(m-Y) \\
&+AT_{t_1,g_c}\left(1_{\{t_1\}}\left(\max\left(m; m+\Delta\mathbb{E}\right)-Y\right) + 1_{\{t_0\}}\left(\min\left(M; M+\Delta\mathbb{E}\right)-Y\right)\right) \\
&+AT_{t_0,g_c}\left(1_{\{t_1\}}\left(\max\left(m; m-\Delta\mathbb{E}\right)-Y\right) + 1_{\{t_0\}}\left(\min\left(M; M-\Delta\mathbb{E}\right)-Y\right)\right).
\end{aligned}$$

Theorem 2.4 ***Partial identification 3***

Under A.1, if $\mathbb{P}(m \leq Y(0) \leq M) = 1$ *with* $(m, M) \in \mathbb{R}^2$, *and* $ATT_{t_1,g_c} = ATT_{t_0,g_c}$,

$$B_-^{''} \leq ATT_{t_1,g_t} \leq B_+^{''},$$

with

$$B_-^{''} = \max\left(\mathbb{E}(Y_{t_1,g_t}|\, D=1) - M; \frac{DID_{Y_-^2}}{\mathbb{P}(D_{t_1,g_t}=1)}\right)$$

and

$$B_+^{''} = \min\left(\mathbb{E}(Y_{t_1,g_t}|\, D=1) - m; \frac{DID_{Y_+^2}}{\mathbb{P}(D_{t_1,g_t}=1)}\right)$$

$B_-^{''}$ *and* $B_+^{''}$ *are sharp.*

As mentioned above, the common trend assumption can be rationalized by the following DGP:

$$Y(d) = f_d(T,U) + g_d(G,U),$$

with $U \perp\!\!\!\perp (T,G)$. In such a framework, $ATT_{t_1,g_c} = ATT_{t_0,g_c}$ can be rationalized under two supplementary assumptions: $f_0(t_1,.) - f_0(t_0,.) = f_1(t_1,.) - f_1(t_0,.)$ and $U \perp\!\!\!\perp T|G=g_c, D=1$. The first assumption means that the effect of time should be the same on $Y(1)$ than on $Y(0)$. The second one means that treated observations in the control group should not be "too different" in period 0 and 1. To assess the credibility of this last hypothesis, one can for instance verify that those two groups of observations do not have very different observable characteristics.

$B_-^{''}$ is obtained as follows. If $ATT_{t_1,g_c} = ATT_{t_0,g_c} = ATT_{g_c}$, (1) rewrites as

$$ATT_{t_1,g_t} = \frac{DID_Y + ATT_{t_0,g_t} \times \mathbb{P}(D_{t_0,g_t}=1) + ATT_{g_c}\left(\mathbb{P}(D_{t_1,g_c}=1) - \mathbb{P}(D_{t_0,g_c}=1)\right)}{\mathbb{P}(D_{t_1,g_t}=1)}. \quad (3)$$

Moreover,

$$\mathbb{E}\left(Y_{t_1,g_c} - M \middle| D = 1\right) \leq ATT_{t_1,g_c} \leq \mathbb{E}\left(Y_{t_1,g_c} - m \middle| D = 1\right)$$

and

$$\mathbb{E}\left(Y_{t_0,g_c} - M \middle| D = 1\right) \leq ATT_{t_0,g_c} \leq \mathbb{E}\left(Y_{t_0,g_c} - m \middle| D = 1\right)$$

implies that

$$\max\left(\mathbb{E}\left(Y_{t_1,g_c} \middle| D = 1\right); \mathbb{E}\left(Y_{t_0,g_c} \middle| D = 1\right)\right) - M \leq ATT_{g_c} \leq \min\left(\mathbb{E}\left(Y_{t_1,g_c} \middle| D = 1\right); \mathbb{E}\left(Y_{t_0,g_c} \middle| D = 1\right)\right) - m.$$

Plugging this last inequality into (3) yields a first lower bound for ATT_{t_1,g_t}. If $\mathbb{P}(D_{t_1,g_c} = 1) \geq \mathbb{P}(D_{t_0,g_c} = 1)$, one should use the left-hand side of the inequality, while if $\mathbb{P}(D_{t_1,g_c} = 1) < \mathbb{P}(D_{t_0,g_c} = 1)$, one should use the right-hand side. This first lower bound can be rewritten as

$$\frac{DID_{Y^2_-}}{\mathbb{P}(D_{t_1,g_t} = 1)}.$$

Finally, since ATT_{t_1,g_t} is also greater than $\mathbb{E}(Y_{t_1,g_t} | D = 1) - M$, B''_- is defined as the maximum of those two lower bounds.

From (3), one can see that the length of $\left[B''_-; B''_+\right]$ essentially depends on

$$\mathbb{P}(D_{t_0,g_t} = 1) + \left|\mathbb{P}(D_{t_1,g_c} = 1) - \mathbb{P}(D_{t_0,g_c} = 1)\right|.$$

Therefore, $\left[B''_-; B''_+\right]$ will be shorter than $[B_-; B_+]$ whose length is proportional to

$$\mathbb{P}(D_{t_0,g_t} = 1) + \mathbb{P}(D_{t_1,g_c} = 1) + \mathbb{P}(D_{t_0,g_c} = 1).^9$$

A situation in which those bounds will be tight is when the natural experiment under consideration is the extension of a policy to a group previously not eligible, and the previously eligible group is used as the control group. Indeed, in such cases $\mathbb{P}(D_{t_0,g_t} = 1) = 0$. Consequently, if the change in the treatment rate from period 0 to 1 in the control group is "small", $\left[B''_-; B''_+\right]$ will be tight. Point identification can even be obtained if $\mathbb{P}(D_{t_1,g_c} = 1) = \mathbb{P}(D_{t_0,g_c} = 1)$. In such instances $\left[B''_-; B''_+\right]$ improves a lot on $[B_-; B_+]$ which will be wide as soon as the percentage of observations treated in the control group is "large".

I illustrate what "small" and "large" mean through another numerical example. As above, I consider a binary outcome Y, such that $DID_Y = 0.05$. I assume that $\mathbb{P}(D_{t_0,g_t} = 1) = 0$ and $\mathbb{P}(D_{t_1,g_t} = 1) = 0.5$: this corresponds to the situation where the treatment group was not eligible for treatment in period 0. I also assume that $\mathbb{P}(D_{t_0,g_c} = 1) = 0.4$: 40% of the control

[9] This change in the size of the identification region is similar to the change happening when using Lee bounds (see Lee (2009) and Horowitz & Manski (1995)) instead of Manski bounds (see Manski (1990)) to deal with missing data.

group was treated in period 0. Finally, I assume that $\mathbb{E}(Y_{t_1,g_c}|\,D=1)=\mathbb{E}(Y_{t_0,g_c}|\,D=1)=0.6$. Under those assumptions, I can write B_-, B_+, $B_-^{''}$ and $B_+^{''}$ as functions of $\mathbb{P}(D_{t_1,g_c}=1)$, the share of treated observations in the control group in period 1. Results are summarized in Table 3.

Table 3: Value of B_-, B_+, $B_-^{''}$ and $B_+^{''}$ according to $\mathbb{P}(D_{t_1,g_c}=1)$

$\mathbb{P}(D_{t_1,g_c}=1)$	30%	32.5%	35%	37.5%	40%	42.5%	45%	47.5%	50%
B_-	-40%	-40%	-40%	-40%	-40%	-40%	-40%	-40%	-40%
B_+	60%	60%	60%	60%	60%	60%	60%	60%	60%
$B_-^{''}$	-2%	1%	4%	7%	10%	8%	6%	4%	2%
$B_+^{''}$	18%	16%	14%	12%	10%	13%	16%	19%	22%

Because of the large shares of treated observations in the control group, B_- and B_+ are not informative: B_- is merely equal to $\mathbb{E}(Y_{t_1,g_t}|\,D=1)-M=0.6-1=-0.4$ and B_+ to $\mathbb{E}(Y_{t_1,g_t}|\,D=1)-m=0.6$. On the contrary, $B_-^{''}$ and $B_+^{''}$ are informative. The length of $\left[B_-^{''};B_+^{''}\right]$ is proportional to $|\mathbb{P}(D_{t_1,g_c}=1)-\mathbb{P}(D_{t_0,g_c}=1)|$. In this particular example, the sign of ATT_{t_1,g_t} is identified as long as the treatment rate does not change by more than 10 percentage points in the control group, which is substantial.

Finally, estimation of $B_-^{''}$ and $B_+^{''}$ requires estimating $\mathbb{P}(D_{t_0,g_c}=1)$, $\mathbb{P}(D_{t_1,g_c}=1)$, $\mathbb{E}(Y_{t_1,g_c}|D=1)$ and $\mathbb{E}(Y_{t_0,g_c}|D=1)$ in a first stage, before conducting the simple OLS regression presented above on Y_-^2 and Y_+^2.

A.1 is a strong identifying assumption which might appear implausible if pre-treatment characteristics that are thought to be associated to the dynamics of the outcome variable are unbalanced between the treatment and the control groups as emphasized in Abadie (2005). In such instances, a conditional common trend assumption might appear more credible. When compliance is perfect, Abadie shows that a standard DID still identifies an ATT under a conditional common trend assumption, except that control group observations should be pre-multiplied by a function w^X, with

$$w^X=\frac{\mathbb{P}(G=g_t|X)}{\mathbb{P}(G=g_c|X)}\times\frac{\mathbb{P}(G=g_t)}{\mathbb{P}(G=g_c)}.$$

Therefore, the control scheme developed by Abadie works by weighting-down observations in the control group for those values of the covariates which are over-represented among the control group, and weighting-up those for those values of the covariates under-represented among the control group, exactly as in a propensity score matching (see Rosenbaum & Rubin, 1983).

All the results presented above can be extended under a conditional common trend assumption. Indeed, I develop fuzzy semi-parametric DID results inspired from Abadie (2005) which allow to control semi-parametrically for covariates. Most of those results merely require to premultiply control group observations by the function w^X and then to conduct the same analysis as above, exactly as in Abadie (2005). Only the extension of Theorem 2.4 is more involved. Therefore, I only present this result here, all the other results are to be found in the appendix.

I introduce two new assumptions taken from Abadie (2005).

A.1' Conditional common trend

With probability one,

$$\mathbb{E}(Y_{t_1,g_t}(0)|X) - \mathbb{E}(Y_{t_0,g_t}(0)|X) = \mathbb{E}(Y_{t_1,g_c}(0)|X) - \mathbb{E}(Y_{t_0,g_c}(0)|X)$$

A.2 Data

The data is either a panel, or consists of two random samples of the same population at dates t_0 and t_1 such that at each date Y, D and G are observed, and covariates X at date t_0 are observed for both samples.

A.2 means that even when pooled cross sections are used in the analysis, covariates should be observed in period 0 both for the period 0 and for the period 1 samples. As discussed below, this might be an issue for time-varying covariates when pooled-crossed sections are used.

I also introduce new notations. For any random variable Z, let

$$DID_Z^X = \mathbb{E}(Z_{t_1,g_t}|X) - \mathbb{E}(Z_{t_0,g_t}|X) - [\mathbb{E}(Z_{t_1,g_c}|X) - \mathbb{E}(Z_{t_0,g_c}|X)]$$

be the DID on Z conditional on X. Let

$$ATT_{i,j}^X = \mathbb{E}(Y_{i,j}(1) - Y_{i,j}(0)|\, D = 1, X),\ \forall (i,j) \in \{t_0; t_1\} \times \{g_c; g_t\}$$

be the effect of the treatment on Y conditional on X among treated observations in the $(i;j)$ cell. For a random variable Z, let

$$WDID_Z = \mathbb{E}(Z_{t_1,g_t}) - \mathbb{E}(Z_{t_0,g_t}) - \left[\mathbb{E}(Z_{t_1,g_c}w^X) - \mathbb{E}(Z_{t_0,g_c}w^X)\right]$$

be a DID on Z in which control group observations are given a weight w^X. Let

$$p^X = \frac{\mathbb{P}(D_{t_0,g_c} = 1|X)}{\mathbb{P}(D_{t_1,g_c} = 1|X)}.$$

p^X is greater than one for control group observations which, conditional on their covariates, have a greater probability of being treated in period 0 than in period 1, and conversely. Let

$$Y_-^3 = Y + (M-Y)AT_{t_0,g_t} + \max\left((M-Y)\left(1-p^X\right);(m-Y)\left(1-p^X\right)\right)AT_{t_1,g_c},$$

$$Y_+^3 = Y + (m-Y)AT_{t_0,g_t} + \min\left((M-Y)\left(1-p^X\right);(m-Y)\left(1-p^X\right)\right)AT_{t_1,g_c},$$

$$Y_-^4 = Y + (M-Y)AT_{t_0,g_t} + \min\left((M-Y)\left(1-\frac{1}{p^X}\right);(m-Y)\left(1-\frac{1}{p^X}\right)\right)AT_{t_0,g_c}$$

and

$$Y_+^4 = Y + (m-Y)AT_{t_0,g_t} + \max\left((M-Y)\left(1-\frac{1}{p^X}\right);(m-Y)\left(1-\frac{1}{p^X}\right)\right)AT_{t_0,g_c}.$$

Before stating the result, I must first state a lemma.

Lemma 2.2 ***Conditional non-identification***

Under A.1',

$$\begin{aligned} DID_Y^X = {} & ATT_{t_1,g_t}^X \times \mathbb{P}(D_{t_1,g_t}=1|X) - ATT_{t_0,g_t}^X \times \mathbb{P}(D_{t_0,g_t}=1|X) \\ & -ATT_{t_1,g_c}^X \times \mathbb{P}(D_{t_1,g_c}=1|X) + ATT_{t_0,g_c}^X \times \mathbb{P}(D_{t_0,g_c}=1|X) \end{aligned} \tag{4}$$

with probability one.

Lemma 2.2 is merely a conditional version of Lemma 2.1. It states that under A.1', DID_Y^X can be rewritten as a weighted DID of four conditionals ATT. From Lemma 2.2, one can show Theorem 2.5.

Theorem 2.5 ***Conditional partial identification***

Under A.1' and A.2, if $\mathbb{P}(m \leq Y(0) \leq M) = 1$ *with* $(m,M) \in \mathbb{R}^2$, $ATT_{t_1,g_c}^X = ATT_{t_0,g_c}^X$ *with probability one, and* $\mathbb{P}(D_{t_1,g_c}=1|X) > 0 \Leftrightarrow \mathbb{P}(D_{t_0,g_c}=1|X) > 0$ *with probability one, then*

$$WB_-^{''} \leq ATT_{t_1,g_t} \leq WB_+^{''},$$

with

$$WB_-^{''} = \max\left(\mathbb{E}(Y_{t_1,g_t}|\,D=1) - M; \frac{WDID_{Y_-^3}}{\mathbb{P}(D_{t_1,g_t}=1)}; \frac{WDID_{Y_-^4}}{\mathbb{P}(D_{t_1,g_t}=1)}\right)$$

and

$$WB_+^{''} = \min\left(\mathbb{E}(Y_{t_1,g_t}|\,D=1) - m; \frac{WDID_{Y_+^3}}{\mathbb{P}(D_{t_1,g_t}=1)}; \frac{WDID_{Y_+^4}}{\mathbb{P}(D_{t_1,g_t}=1)}\right)$$

$WB_{-}^{''}$ is obtained as follows. A mere integration over the distribution of X yields

$$ATT_{t_1,g_t} = \int ATT^X_{t_1,g_t} dP(X|D_{t_1,g_t} = 1). \tag{5}$$

From Lemma 2.2 one gets that

$$ATT^X_{t_1,g_t} = \frac{DID^X_Y + ATT^X_{t_0,g_t} \times \mathbb{P}(D_{t_0,g_t} = 1|X) + ATT^X_{t_1,g_c} \times \mathbb{P}(D_{t_1,g_c} = 1|X) - ATT^X_{t_0,g_c} \times \mathbb{P}(D_{t_0,g_c} = 1|X)}{\mathbb{P}(D_{t_1,g_t} = 1|X)}. \tag{6}$$

If $ATT^X_{t_1,g_c} = ATT^X_{t_0,g_c}$, (6) rewrites both as

$$ATT^X_{t_1,g_t} = \frac{DID^X_Y + ATT^X_{t_0,g_t} \times \mathbb{P}(D_{t_0,g_t} = 1|X) + ATT^X_{t_1,g_c} \times (\mathbb{P}(D_{t_1,g_c} = 1|X) - \mathbb{P}(D_{t_0,g_c} = 1|X))}{\mathbb{P}(D_{t_1,g_t} = 1|X)} \tag{7}$$

and

$$ATT^X_{t_1,g_t} = \frac{DID^X_Y + ATT^X_{t_0,g_t} \times \mathbb{P}(D_{t_0,g_t} = 1|X) + ATT^X_{t_0,g_c} \times (\mathbb{P}(D_{t_1,g_c} = 1|X) - \mathbb{P}(D_{t_0,g_c} = 1|X))}{\mathbb{P}(D_{t_1,g_t} = 1|X)}. \tag{8}$$

Therefore, to derive a first lower bound for ATT_{t_1,g_t}, I start bounding $ATT^X_{t_1,g_c}$ by $\mathbb{E}\left(Y_{t_1,g_c} - M|\, D = 1, X\right)$ when $\mathbb{P}(D_{t_1,g_c} = 1|X) \geq \mathbb{P}(D_{t_0,g_c} = 1|X)$ and by $\mathbb{E}\left(Y_{t_1,g_c} - m|\, D = 1, X\right)$ when $\mathbb{P}(D_{t_1,g_c} = 1|X) < \mathbb{P}(D_{t_0,g_c} = 1|X)$ in (7). This yields a bound for $ATT^X_{t_1,g_t}$ which I plug into (5). Then I use Bayes rule to write this bound as a quantity which can be easily estimated from the sample since it is equal to

$$\frac{WDID_{Y_{-}^3}}{\mathbb{P}(D_{t_1,g_t} = 1)}.$$

Following the same steps, I derive a second lower bound for ATT_{t_1,g_t} using (8), which is equal to

$$\frac{WDID_{Y_{-}^4}}{\mathbb{P}(D_{t_1,g_t} = 1)}.$$

Finally, since ATT_{t_1,g_t} is also greater than $\mathbb{E}(Y_{t_1,g_t}|\, D = 1) - M$, $WB_{-}^{''}$ is defined as the maximum of those three lower bounds.

Those fuzzy semi-parametric DID results improve on the fuzzy DID results on two dimensions. First, conditional common trend is more credible than common trend in many contexts as emphasized above. Moreover, Theorem 2.5 no longer requires that $ATT_{t_1,g_c} = ATT_{t_0,g_c}$ but that $ATT^X_{t_1,g_c} = ATT^X_{t_0,g_c}$. This is probably a more credible assumption, all the more so if observable characteristics of treated observations in the control group change between period 0 and 1.

The main limitation of those fuzzy semi-parametric DID results is that they require that covariates are observed in period 0 both for the period 0 and for the period 1 samples, exactly as for Abadie's semi-parametric DID results. This might raise a data availability issue for time-varying covariates when pooled-crossed sections are used. In such instances, since all

the parameters identified above can be estimated through simple linear regressions, another option to control for covariates could merely be to include them in those regressions. For instance, one could run the following regression

$$Y_{-}^{0} = \alpha + \beta T + \gamma G + \delta X + \mu XT + \theta TG + u, \tag{9}$$

to estimate $DID_{Y_{-}^{0}}$ taking into account the effect of covariates on the dynamics of the outcome. The main limit of this second approach is that, as emphasized in Meyer (1995), introducing covariates in this linear fashion may not be appropriate if the treatment has different effects for different groups in the population.

3 Inference

The objective of this section is to build up confidence intervals (CI) for ATT_{t_1,g_t} based upon the results of section 2. Deriving such confidence intervals from the point identification results is straightforward: one can use the delta method to derive the asymptotic distribution of $\frac{\widehat{DID}_Y}{\widehat{DID}_D}$. Using partial identification results proves less straightforward. An extensive literature on confidence intervals for partially identified parameters developed recently. Some solutions have been proposed by Imbens & Manski (2004) and Stoye (2009) but they do not apply here. Indeed, they require uniform asymptotic normality of the estimators. As shown in Hirano & Porter (2009), there exists no regular estimators of parameters defined as non continuously differentiable functional of the data distribution. B_{-} and B_{+}, just as all the other bounds presented in section 2, are not continuously differentiable functionals of the data distribution. Indeed,

$$B_{-} = \max\left(\mathbb{E}(Y_{t_1,g_t} |\, D = 1) - M; \frac{DID_{Y_{-}^{0}}}{\mathbb{P}(D_{t_1,g_t} = 1)}\right)$$

for instance.

Therefore, I use a recent method from the moment inequality literature, developed by Andrews & Soares (2010) which applies to my setting and does not require this regularity condition.[10]

Let $\theta_0 = ATT_{t_1,g_t}$ be my parameter of interest and F_0 be the true distribution of the data.

[10] Chernozhukov et al. (2011) does not apply here because my bounds are not intersection bounds.

When $m \leq Y(0) \leq M$, Theorem 2.2 implies that θ_0 verifies four inequalities:

$$\begin{aligned} \theta_0 - B^1 &\geq 0 \\ \theta_0 - B^2 &\geq 0 \\ B^3 - \theta_0 &\geq 0 \\ B^4 - \theta_0 &\geq 0, \end{aligned} \tag{10}$$

with $B^1 = \frac{DID_{Y^0_-}}{\mathbb{P}(D_{t_1,g_t}=1)}$, $B^2 = \mathbb{E}(Y_{t_1,g_t} - M |\, D = 1)$, $B^3 = \frac{DID_{Y^0_+}}{\mathbb{P}(D_{t_1,g_t}=1)}$ and $B^4 = \mathbb{E}(Y_{t_1,g_t} - m |\, D = 1)$. The set of all (θ, F) compatible with those moment inequalities is denoted $\mathcal{F}$.

Let

$$T_n(\theta) = \sum_{j=1}^{2} [\sqrt{n} \frac{\theta - \widehat{B^j}}{\widehat{\sigma_j}}]_-^2 + \sum_{j=3}^{4} [\sqrt{n} \frac{\widehat{B^j} - \theta}{\widehat{\sigma_j}}]_-^2,$$

where $[x]_- = x$ if $x < 0$ and 0 otherwise and where σ_j is the asymptotic variance of $\sqrt{n}\left(B^j - \widehat{B^j}\right)$. Let D_j be an indicator variable equal to 1 when $\sqrt{n}\frac{\theta_0 - \widehat{B^j}}{\widehat{\sigma_j}} \leq \sqrt{2\ln(\ln(n))}$ for $j = 1$ or 2, and when $\sqrt{n}\frac{\widehat{B^j} - \theta_0}{\widehat{\sigma_j}} \leq \sqrt{2\ln(\ln(n))}$ for $j = 3$ or 4. Let CS_n be the confidence interval obtained inverting $T_n(\theta)$, using as critical values $c_{1-\alpha}(\theta)$ the $(1-\alpha)^{th}$ quantile of the distribution of

$$\sum_{1 \leq j \leq 4 / D_j = 1} [N^j]_-^2,$$

where $\left(N^1, N^2, N^3, N^4\right)'$ is a vector of $\mathcal{N}(0,1)$ random variables with a variance Ω equal to the asymptotic variance of

$$\sqrt{n}\left(\frac{B^1 - \widehat{B^1}}{\widehat{\sigma_1}}, \frac{B^2 - \widehat{B^2}}{\widehat{\sigma_2}}, \frac{\widehat{B^3} - B^3}{\widehat{\sigma_3}}, \frac{\widehat{B^4} - B^4}{\widehat{\sigma_4}}\right)'.$$

Formally,

$$CS_n = \{\theta \;/\; T_n(\theta) \leq c_{1-\alpha}(\theta)\}.$$

Theorem 3.1 ***Inference***

When $m \leq Y(0) \leq M$, if $m' \leq Y(1) \leq M'$ with $(m', M') \in \mathbb{R}^2$ and if there exists $\epsilon > 0$ such that $\mathbb{P}(T = i, G = j) \geq \epsilon$, for every $(i,j) \in \{t_0; t_1\} \times \{g_c; g_t\}$, then CS_n is uniformly valid and not conservative:

$$\liminf_{n \to +\infty} \inf_{(\theta,F) \in \mathcal{F}} \mathbb{P}_F\left(T_n(\theta) \leq c_{1-\alpha}(\theta)\right) = 1 - \alpha.$$

The proof of Theorem 3.1 essentially amounts to show that assumptions of the inequality model developed by Andrews & Soares (2010) hold here. In order to do so, I take a few more

assumptions than in Theorem 2.2. The assumption that $Y(1)$ also is bounded is not very restrictive: it is hard to envision real life situations where $Y(0)$ is bounded but not $Y(1)$. The assumption that there exists $\epsilon > 0$ such that $\mathbb{P}(T = i, G = j) \geq \epsilon$ is just to rule out degenerate probability distributions. Since all the inequalities are linear in θ_0, inverting the test is fairly simple.

The intuition of this result is as follows. For $j = 1$ or 2, if $\theta_0 > B^j$,

$$\lim_{n \to +\infty} \sqrt{n} \frac{\theta_0 - \widehat{B^j}}{\widehat{\sigma_j}} = +\infty$$

with probability one. Similarly, for $j = 3$ or 4, if $B^j > \theta_0$,

$$\lim_{n \to +\infty} \sqrt{n} \frac{\widehat{B^j} - \theta_0}{\widehat{\sigma_j}} = +\infty.$$

Therefore, $T_n(\theta_0)$ converges in distribution towards

$$\sum_{1 \leq j \leq 4 / \theta_0 = B^j} [N^j]_-^2.$$

Consequently, the limiting distribution of the test statistic and the critical values to be used when inverting it would be known if we knew which inequalities are actually equalities. To solve this problem, Andrews & Soares (2010) introduce the following decision rule: they consider an inequality as an equality when $\sqrt{n} \frac{\theta_0 - \widehat{B^j}}{\widehat{\sigma_j}} \leq \sqrt{2 \ln(\ln(n))}$ (or when $\sqrt{n} \frac{\theta_0 - \widehat{B^j}}{\widehat{\sigma_j}} \leq \sqrt{2 \ln(\ln(n))}$ for $j = 3$ or 4) , i.e. when $\sqrt{n} \frac{\theta_0 - \widehat{B^j}}{\widehat{\sigma_j}}$ is not "too large". The reason why the resulting CI is uniformly valid is that when $\theta_0 = B^j$, $\sqrt{n} \frac{\theta_0 - \widehat{B^j}}{\widehat{\sigma_j}}$ converges to a $\mathcal{N}(0,1)$ random variable. Since $\sqrt{2 \ln(\ln(n))}$ converges to $+\infty$, the probability to wrongly consider an equality as an inequality vanishes to 0.

Finally, Theorems 2.3 and 2.4 can also be written as moment inequality models. Therefore, it is also possible to derive CI for θ_0 from $\widehat{B'_-}$ and $\widehat{B'_+}$, or from $\widehat{B''_-}$ and $\widehat{B''_+}$, using the method of Andrews & Soares (2010). Since Theorem 2.3 does not require that Y is bounded, when $\widehat{B'_-}$ and $\widehat{B'_+}$ are used to construct a CI for θ_0, one will have to add the technical assumption that $\mathbb{E}_F\left(|Y(d)|^{2+\delta}\right) < K$ for $d = 0$ or 1, where K is a constant and $\delta > 0$.

4 The impact of varenicline on smoking cessation.

4.1 Data and methods

The national data base of French smoking cessation clinics started in 2001. 59 clinics recorded at least one patient per year in 2006 and 2007. During patients first visit, smoking status

is evaluated according to daily cigarettes smoked and a measure of expired carbon monoxide (CO) which is a biomarker for recent tobacco use. At the end of this baseline visit, treatments may be prescribed to patients (nicotine replacement therapies...). Follow-up visits are offered during which CO measures are usually made to check whether the patient remained abstinent.

Varenicline is a pharmacotherapy for smoking cessation support which was made available to these clinics in February 2007. The kernel density estimate of the rate of prescription of varenicline per clinic is shown in Figure 1. It is bimodal, with a first peak at very low rates of prescription, and a second smaller peak around 35-40%. In 15 clinics, less than 3% of all patients consulted have been prescribed varenicline during the year following its release. In 13 clinics, more than 20% of patients have received this new drug. I exploit this to estimate the impact of varenicline on smoking cessation. The control group is made up of patients registered by the 15 "below 3% prescription rate" clinics, hereafter referred to as the control clinics. The treatment group consists of patients recorded by "above 20% prescription rate" clinics, hereafter referred to as treatment clinics. Period 0 goes from February 2006 to January 2007, and period 1 from February 2007 to January 2008.

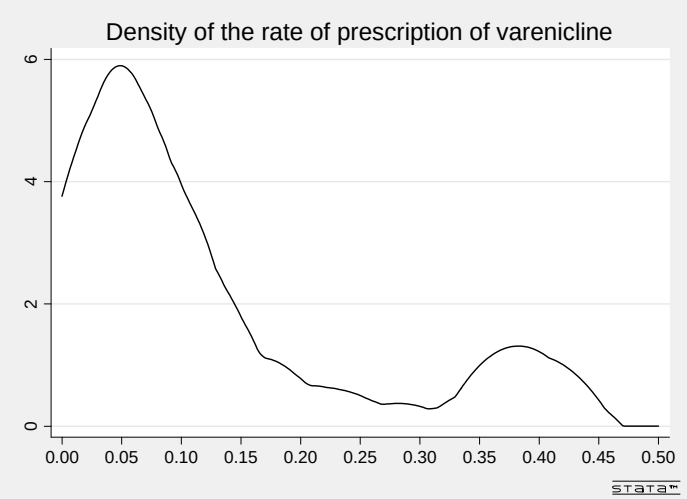

Figure 1: Density of the prescription rate of Varenicline across clinics.

8 581 patients consulted those 28 clinics over period 0 and 1. Because many patients never came back for follow-up visits, there are only 5 299 patients (62% of the initial sample) for whom follow-up CO measures are available. I exclude patients for whom no such measures are available from the analysis. Among remaining patients, the outcome variable is a dummy equal to 1 for patients whose last follow-up CO determination was inferior or equal to 5 parts per million (ppm).

4.2 Results

In Table 4, I provide descriptive statistics on patients per group of clinics and per period of time. Patients consulted in those cessation clinics are middle-aged, rather educated and the majority of them are employed. They are very heavy smokers since they smoke more than 21.6 cigarettes per day on average, which corresponds to the 90th percentile in the French distribution of smokers (Beck et al. (2007)).

Table 4: Descriptive Statistics

	Treatment Clinics			**Control Clinics**		
	2006	2007	P-value	2006	2007	P-value
Patients' characteristics						
% Males	47.9%	47.9%	0.95	48.5%	50.4%	0.30
Age	44.6	43.7	0.08	44.0	44.3	0.52
% employed	65.3%	68.3%	0.11	65.3%	69.8%	0.01
% without degree	19.2%	21.0%	0.25	14.2%	14.1%	0.98
Daily cigarettes smoked	21.7	21.9	0.6	22.1	20.9	<0.01
Treatment prescribed						
% prescribed nicotine patch	75.0%	45.5%	<0.001	45.9%	49.7%	0.05
% prescribed varenicline	0.01%	38.2%	<0.001	0%	1.6%	<0.001
% of successful quits	53.7%	56.9%	0.11	46.6%	41.6%	<0.01
N	1 195	1 303		1 300	1 501	

In period 0, the prescription rate of varenicline was equal to 0% in control clinics and to 0.01% in treatment clinics.[11] In period 1, it rose to 1.6% in control clinics and to 38.2% in treatment clinics. This sharp rise in varenicline prescription in treatment clinics entailed a strong decrease in the prescription of other treatments such as nicotine patch. Finally, from period 0 to 1, abstinence rate increased (from 53.7% to 56.9%) in treatment clinics, whereas it decreased (from 46.6% to 41.6%) in control clinics.

From Theorem 2.2, I compute that $\widehat{B_-} = 19.1\%$ and $\widehat{B_+} = 24.5\%$. Using Theorem 3.1, I find that $[0.074; 0.364]$ is a 95% confidence interval for ATT_{t_1,g_t}. Some covariates are not balanced across treatment and control patients. For instance, patients in treatment clinics are less educated. If those covariates also have an impact on the dynamic of the outcome, the common trend assumption might be violated, hence the need to control for them. Many covariates available in the data are time varying, such as daily cigarettes smoked. Therefore, I cannot use the semi-parametric fuzzy DID results of section 2. Instead, I merely estimate $DID_{Y_-^0}$ including controls (sex, age, employment status, daily cigarettes smoked and CO at

[11] Varenicline was prescribed to 6 patients recorded in the last week of January 2007, that is to say a few days before its release.

baseline) into the regression as in equation (9). This results in a slight change of $\widehat{B_-}$ which increases to 22.9%. $\widehat{B_+}$ also slightly increases.

4.3 Robustness checks

The main assumption needed to identify $[B_-; B_+]$ is the common trend assumption. To indirectly assess its validity, I use the fact that I have several years of data available and I compute placebo DID from 2003 to 2008. They are displayed in Table 5 along with their P-values. Only the 2006-2007 DID is significant. This means that 2006 and 2007 are the only two years between which the cessation rate did not follow parallel trends in the two groups of clinics. Between 2003 and 2004, 2004 and 2005, 2005 and 2006, and 2007 and 2008, trends in cessation rates were not significantly different in the two groups of clinics. The common trend assumption states that the cessation rate would have followed parallel trends between 2006 and 2007 if varenicline had not been released. The fact that before and after the release of varenicline, cessation rate indeed followed roughly parallel trends in the two groups of clinics gives some credit to that assumption.

Attrition seems orthogonal to the interaction of period 1 and treatment clinics, since the DID computed on the percentage of patients included is low and insignificant (+2.2%, P-value = 0.62). Therefore, estimates do not seem contaminated by attrition bias.

Table 5: Placebo DID

	Placebo DID	P-value	N
On cessation rate			
2003-2004	0.045	0.34	1 580
2004-2005	0.032	0.53	2 499
2005-2006	0.042	0.31	4 136
2006-2007	0.082	0.01	5 299
2007-2008	-0.043	0.46	4 400
On attrition rate			
2006-2007	0.022	0.62	8 581

Note: Standard errors are clustered at the clinic level.

Finally, one might worry about the arbitrariness of the definition of my treatment and control groups which is not based on some objective characteristic of cessation clinics. I investigate the sensitivity of the results to the 3%-20% rule as a robustness check. Here, since there are very few always takers, $B_- = B^1$. I run the same analysis with 9 different pairs of thresholds, which makes the sample of clinics included vary from 18 to 38. As shown in Table 6, I always get $\widehat{B^1} \geq 0$, with 5 P-values lower than 0.10. Results are less significant when the

threshold used for control clinics is a prescription rate of 4%. Firstly, this is not necessarily surprising since the bounds are less tight when there are more always takers. Moreover, those results become significant again when controls are included.

Table 6: P-value of $\widehat{B^1}$ according to inclusion thresholds

	Treatment threshold: 15%	Treatment threshold: 20%	Treatment threshold: 25%
Control threshold: 2%	0.10	0.08	0.14
Control threshold: 3%	0.03	0.02	0.06
Control threshold: 4%	0.20	0.15	0.27

Note: Standard errors are clustered at the clinic level.

4.4 Why do treatment and control clinics have different prescription rates ?

Finally, I tentatively investigate where the difference in varenicline prescription rates across clinics comes from. It might merely come from the fact patients coming to those two groups of clinics are different and need different drugs. A simple probit regression of varenicline prescription indicates that it is positively related to patients employment status, daily cigarettes smoked and addiction levels. However, treatment and control patients consulted in period 1 significantly differ only on cigarettes smoked (1 more cigarette smoked per day in treatment clinics).

A second hypothesis is that professionals working in those clinics differ, either in terms of occupation, qualifications or beliefs about effective ways of accompanying smoking cessation. Information on professionals working in smoking cessations clinics is available for only 7 clinics (4 treatment and 3 control) out of the 28 included in the analysis. Still, the 4 treatment clinics recorded 1 612 patients over period 0 and 1, that is to say 64.5% of the "treatment" sample, and the 3 control clinics recorded 1 828 patients, that is to say 65.3% of the control sample. It appears that within this subsample of the total population, treatment patients had a higher probability of being consulted by a doctor (76% against 47%). On the contrary, they had a lower probability of being consulted by a psychologist (3% against 18%) or by someone trained to behavioral and cognitive therapies (4% against 48%). Finally, treatment clinics consulted 193 new patients per full time working professional in 2007, against only 75 in control clinics.

Contrarily to nicotine replacement therapies, varenicline must be prescribed by a doctor. The sharp difference in prescription rates across the two groups of clinics might therefore come from the lower proportion of doctors in control clinics. But patients consulted in those clinics still had a 47% probability of being consulted by a doctor and only 1.6% were prescribed

varenicline. A complementary explanation is that there might be two approaches to smoking cessation among professionals. The first approach, which seems more prominent in treatment clinics, puts the emphasis on providing patients pharmacotherapies to reduce the symptoms of withdrawal. The second approach, which seems more prominent in control clinics, lays more the emphasis on giving them intensive psychological support, hence the higher share of professionals trained to behavioral and cognitive therapies, the lower number of patients consulted per professional and the lower prescription rate of nicotine patches previous to the release of varenicline.

5 Conclusion

This paper develops new results to identify treatment effects when the treatment rate increases more in one group between two periods. The parameters identified will be informative in applications with no or few always takers, as well as in applications where the treatment rate does not change much between the two periods in the control group. My results essentially hold under a common trend assumption on the outcome, whereas the IV-DID result commonly invoked in such settings holds under two common trends assumptions on the outcome and on the treatment, and under a monotonicity assumption. I give examples of applications in which it might be appealing to dispose of the common trend assumption on D while maintaining it on Y.

References

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

Andrews, D. W. K. & Guggenberger, P. (2010), 'Asymptotic size and a problem with subsampling and with the m out of n bootstrap', *Econometric Theory* **26**(02), 426–468.

Andrews, D. W. K. & Soares, G. (2010), 'Inference for parameters defined by moment inequalities using generalized moment selection', *Econometrica* **78**(1), 119–157.

Angrist, J. D. & Evans, W. N. (1998), 'Children and their parents' labor supply: Evidence from exogenous variation in family size', *American Economic Review* **88**(3), 450–77.

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), 431–442.

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

Bach, L. (2010), Are small-and-medium-sized firms really credit constrained ? evidence from a french targeted credit programme, Working paper.

Battistin, E. & Rettore, E. (2008), 'Ineligibles and eligible non-participants as a double comparison group in regression-discontinuity designs', *Journal of Econometrics* **142**(2), 715–730.

Chernozhukov, V., Lee, S. S. & Rosen, A. (2011), Intersection bounds: estimation and inference, Technical report, Centre for Microdata Methods and Practice, Institute for Fiscal Studies.

de Chaisemartin, C. (2011), Instrumented differences in differences, Working paper.

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.

Eissa, N. & Liebman, J. B. (1996), 'Labor supply response to the earned income tax credit', *The Quarterly Journal of Economics* **111**(2), 605–637.

Evans, W. N. & Ringel, J. S. (1999), 'Can higher cigarette taxes improve birth outcomes?', *Journal of Public Economics* **72**(1), 135–154.

Field, E. (2005), 'Property rights and investment in urban slums', *Journal of the European Economic Association* **3**(2-3), 279–290.

Hirano, K. & Porter, J. (2009), Impossibility results for nondifferentiable functionals, Mpra paper, University Library of Munich, Germany.

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

Imbens, G. W. & Angrist, J. D. (1994), 'Identification and estimation of local average treatment effects', *Econometrica* **62**(2), 467–475.

Imbens, G. W. & Manski, C. F. (2004), 'Confidence intervals for partially identified parameters', *Econometrica* **72**(6), 1845–1857.

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.

Manski, C. F. (1997), 'Monotone treatment response', *Econometrica* **65**(6), 1311–1334.

Meyer, B. D. (1995), 'Natural and quasi-experiments in economics', *Journal of Business & Economic Statistics* **13**(2), 151–61.

Rosenbaum, P. R. & Rubin, D. B. (1983), 'The central role of the propensity score in observational studies for causal effects', *Biometrika* **70**(1), pp. 41–55.

Rubin, D. (1974), 'Estimating causal effects of treatments in randomized and nonrandomized studies', *Journal of Educational Psychology* **66**(5).

Stoye, J. (2009), 'More on confidence intervals for partially identified parameters', *Econometrica* **77**(4), 1299–1315.

Appendix A: Identification under conditional common trend

Theorem 5.1 ***Conditional identification***

1. *Under A.1' and A.2, in the no always takers special case,*

$$ATT_{t_1,g_t} = \frac{WDID_Y}{\mathbb{P}(D_{t_1,g_t}=1)}$$

2. *Under A.1' and A.2, if* $\mathbb{P}(m \leq Y(0) \leq M) = 1$ *with* $(m, M) \in \mathbb{R}^2$,

$$WB_- \leq ATT_{t_1,g_t} \leq WB_+,$$

with

$$WB_- = \max\left(\mathbb{E}(Y_{t_1,g_t}|\, D=1) - M; \frac{WDID_{Y_-^0}}{\mathbb{P}(D_{t_1,g_t}=1)}\right)$$

and

$$WB_+ = \min\left(\mathbb{E}(Y_{t_1,g_t}|\, D=1) - m; \frac{WDID_{Y_+^0}}{\mathbb{P}(D_{t_1,g_t}=1)}\right).$$

3. *Under A.1' and A.2, if* $\mathbb{P}(m \leq Y(0) \leq Y(1)) = 1$ *with* $m \in \mathbb{R}$,

$$WB_-^{'} \leq ATT_{t_1,g_t} \leq WB_+^{'},$$

with

$$WB_-^{'} = \max\left(0; \frac{WDID_{Y_-^1}}{\mathbb{P}(D_{t_1,g_t}=1)}\right)$$

and

$$WB_+^{'} = \min\left(\mathbb{E}(Y_{t_1,g_t}|\, D=1) - m; \frac{WDID_{Y_+^1}}{\mathbb{P}(D_{t_1,g_t}=1)}\right).$$

Points 1 to 3 of Theorem 5.1 are obtained as follows. Plugging (6) into (5) and using Bayes rule yields

$$ATT_{t_1,g_t} = \frac{WDID_Y + ATT_{t_0,g_t} \times \mathbb{P}(D_{t_0,g_t}=1) + WATT_{t_1,g_c} \times \mathbb{P}(D_{t_1,g_c}=1) - WATT_{t_0,g_c} \times \mathbb{P}(D_{t_0,g_c}=1)}{\mathbb{P}(D_{t_1,g_t}=1)}, \quad (11)$$

where $WATT_{i,g_c} = \mathbb{E}\left((Y_{i,j}(1) - Y_{i,j}(0))\, w^X |\, D=1\right)$, for $i \in \{t_0; t_1\}$, are weighted ATTs in the two control group cells, in which observations are given a weight w^X. In the no always takers case, $\mathbb{P}(D_{t_0,g_t}=1) = \mathbb{P}(D_{t_1,g_c}=1) = \mathbb{P}(D_{t_0,g_c}=1) = 0$, hence the first claim. Then, if $m \leq Y(0) \leq M$, ATT_{t_0,g_t}, $WATT_{t_1,g_c}$ and $WATT_{t_0,g_c}$ can all be bounded, hence the second claim. Finally, if $m \leq Y(0) \leq Y(1)$, I can use this inequality to derive new bounds for ATT_{t_0,g_t}, $WATT_{t_1,g_c}$ and $WATT_{t_0,g_c}$, hence the third claim.

Appendix B: Proofs

Proof of Lemma 2.1:

$$Y = Y(1) \times D + Y(0) \times (1 - D) = (Y(1) - Y(0)) \times D + Y(0).$$

Therefore,

$$\begin{aligned} DID_Y = &\mathbb{E}\left[(Y_{t_1,g_t}(1) - Y_{t_1,g_t}(0))D\right] - \mathbb{E}\left[(Y_{t_0,g_t}(1) - Y_{t_0,g_t}(0))D\right] \\ &-\mathbb{E}\left[(Y_{t_1,g_c}(1) - Y_{t_1,g_c}(0))D\right] + \mathbb{E}\left[(Y_{t_0,g_c}(1) - Y_{t_0,g_c}(0))D\right] \\ &+\mathbb{E}(Y_{t_1,g_t}(0)) - \mathbb{E}(Y_{t_0,g_t}(0)) - \mathbb{E}(Y_{t_1,g_c}(0)) + \mathbb{E}(Y_{t_0,g_c}(0)). \end{aligned}$$

Under A.1,

$$\mathbb{E}(Y_{t_1,g_t}(0)) - \mathbb{E}(Y_{t_0,g_t}(0)) - \mathbb{E}(Y_{t_1,g_c}(0)) + \mathbb{E}(Y_{t_0,g_c}(0)) = 0.$$

Thus,

$$\begin{aligned} DID_Y = &\mathbb{E}(Y_{t_1,g_t}(1) - Y_{t_1,g_t}(0)|\, D = 1) \times \mathbb{P}(D_{t_1,g_t} = 1) \\ &-\mathbb{E}(Y_{t_0,g_t}(1) - Y_{t_0,g_t}(0)|\, D = 1) \times \mathbb{P}(D_{t_0,g_t} = 1) \\ &-\mathbb{E}(Y_{t_1,g_c}(1) - Y_{t_1,g_c}(0)|\, D = 1) \times \mathbb{P}(D_{t_1,g_c} = 1) \\ &+\mathbb{E}(Y_{t_0,g_c}(1) - Y_{t_0,g_c}(0)|\, D = 1) \times \mathbb{P}(D_{t_0,g_c} = 1), \end{aligned}$$

hence the result.

QED.

Proof of Theorem 2.1:

Proof of 1)

In the "no always takers" special case, $\mathbb{P}(D_{t_0,g_t} = 1)$, $\mathbb{P}(D_{t_1,g_c} = 1)$ and $\mathbb{P}(D_{t_0,g_c} = 1)$ are all equal to 0. Therefore, (1) can be rewritten as

$$DID_Y = ATT_{t_1,g_t} \times \mathbb{P}(D_{t_1,g_t} = 1)$$

hence the result.

Proof of 2)

If $\forall (i,j) \in \{t_0; t_1\} \times \{g_c; g_t\}$, $ATT_{i,j} = ATT_{t_1,g_t}$, then, (1) can be rewritten as

$$DID_Y = ATT_{t_1,g_t} \times DID_D,$$

hence the result.

QED.

Proof of Theorem 2.2:

Let

$$A = \mathbb{E}(Y_{t_0,g_t}(0)|\, D = 1) \times \mathbb{P}(D_{t_0,g_t} = 1) + \mathbb{E}(Y_{t_1,g_c}(0)|\, D = 1) \times \mathbb{P}(D_{t_1,g_c} = 1) - \mathbb{E}(Y_{t_0,g_c}(0)|\, D = 1) \times \mathbb{P}(D_{t_0,g_c} = 1).$$

This is the only quantity appearing in (2) which cannot be estimated from the sample and therefore needs to be bounded.
Since $m \leq Y(0) \leq M$,

$$A_1^- \leq A \leq A_1^+,$$

with

$$A_1^- = m \times \mathbb{P}(D_{t_0,g_t} = 1) + m \times \mathbb{P}(D_{t_1,g_c} = 1) - M \times \mathbb{P}(D_{t_0,g_c} = 1)$$

and

$$A_1^+ = M \times \mathbb{P}(D_{t_0,g_t} = 1) + M \times \mathbb{P}(D_{t_1,g_c} = 1) - m \times \mathbb{P}(D_{t_0,g_c} = 1).$$

For bounds to be sharp, a DGP attaining them should also verify the common trend assumption, which implies:

$$\begin{aligned}
0 = &\mathbb{E}(Y_{t_1,g_t}(0)|\, D=1) \times \mathbb{P}(D_{t_1,g_t}=1) + \mathbb{E}(Y_{t_1,g_t}|\, D=0) \times (1 - \mathbb{P}(D_{t_1,g_t}=1)) \\
&-\mathbb{E}(Y_{t_0,g_t}(0)|\, D=1) \times \mathbb{P}(D_{t_0,g_t}=1) - \mathbb{E}(Y_{t_0,g_t}|\, D=0) \times (1 - \mathbb{P}(D_{t_0,g_t}=1)) \\
&-\mathbb{E}(Y_{t_1,g_c}(0)|\, D=1) \times \mathbb{P}(D_{t_1,g_c}=1) - \mathbb{E}(Y_{t_1,g_c}|\, D=0) \times (1 - \mathbb{P}(D_{t_1,g_c}=1)) \\
&+\mathbb{E}(Y_{t_0,g_c}(0)|\, D=1) \times \mathbb{P}(D_{t_0,g_c}=1) + \mathbb{E}(Y_{t_0,g_c}|\, D=0) \times (1 - \mathbb{P}(D_{t_0,g_c}=1)).
\end{aligned}$$

The only quantity in this equation which is both unobserved and does not enter into (2) is $\mathbb{E}(Y_{t_1,g_t}(0)|\, D=1)$. For common trend to hold, it should be equal to

$$\begin{aligned}
&\tfrac{1}{\mathbb{P}(D_{t_1,g_t}=1)} \left(A + \mathbb{E}(Y_{t_0,g_t}|\, D=0) \times (1 - \mathbb{P}(D_{t_0,g_t}=1)) + \mathbb{E}(Y_{t_1,g_c}|\, D=0) \times (1 - \mathbb{P}(D_{t_1,g_c}=1))\right) \\
&-\tfrac{1}{\mathbb{P}(D_{t_1,g_t}=1)} \left(\mathbb{E}(Y_{t_1,g_t}|\, D=0) \times (1 - \mathbb{P}(D_{t_1,g_t}=1)) + \mathbb{E}(Y_{t_0,g_c}|\, D=0) \times (1 - \mathbb{P}(D_{t_0,g_c}=1))\right).
\end{aligned}$$

Since $m \leq \mathbb{E}(Y_{t_1,g_t}(0)|\, D=1) \leq M$, this implies that

$$A_2^- \leq A \leq A_2^+,$$

with

$$\begin{aligned}
A_2^- = &m \times \mathbb{P}(D_{t_1,g_t}=1) - \mathbb{E}(Y_{t_0,g_t}|\, D=0) \times (1 - \mathbb{P}(D_{t_0,g_t}=1)) \\
&-\mathbb{E}(Y_{t_1,g_c}|\, D=0) \times (1 - \mathbb{P}(D_{t_1,g_c}=1)) + \mathbb{E}(Y_{t_1,g_t}|\, D=0) \times (1 - \mathbb{P}(D_{t_1,g_t}=1)) \\
&+\mathbb{E}(Y_{t_0,g_c}|\, D=0) \times (1 - \mathbb{P}(D_{t_0,g_c}=1)).
\end{aligned}$$

and

$$\begin{aligned}
A_2^+ = &M \times \mathbb{P}(D_{t_1,g_t}=1) - \mathbb{E}(Y_{t_0,g_t}|\, D=0) \times (1 - \mathbb{P}(D_{t_0,g_t}=1)) \\
&-\mathbb{E}(Y_{t_1,g_c}|\, D=0) \times (1 - \mathbb{P}(D_{t_1,g_c}=1)) + \mathbb{E}(Y_{t_1,g_t}|\, D=0) \times (1 - \mathbb{P}(D_{t_1,g_t}=1)) \\
&+\mathbb{E}(Y_{t_0,g_c}|\, D=0) \times (1 - \mathbb{P}(D_{t_0,g_c}=1)).
\end{aligned}$$

Consequently,

$$\max(A_1^-; A_2^-) \leq A \leq \min(A_1^+; A_2^+). \tag{12}$$

Combining (2) and (12) and rearranging yields B_- and B_+, which are sharp by construction. **QED.**

Proof of Theorem 2.3:
Since $m \leq Y(0) \leq Y(1)$,

$$A_3^- \leq A \leq A_3^+,$$

with

$$A_3^- = m \times \mathbb{P}(D_{t_0,g_t} = 1) + m \times \mathbb{P}(D_{t_1,g_c} = 1) - \mathbb{E}(Y_{t_0,g_c} | \, D = 1) \times \mathbb{P}(D_{t_0,g_c} = 1)$$

and

$$A_3^+ = \mathbb{E}(Y_{t_0,g_t} | \, D = 1) \times \mathbb{P}(D_{t_0,g_t} = 1) + \mathbb{E}(Y_{t_1,g_c} | \, D = 1) \times \mathbb{P}(D_{t_1,g_c} = 1) - m \times \mathbb{P}(D_{t_0,g_c} = 1).$$

For bounds to be sharp, a DGP attaining them should also verify the common trend assumption. Given that $m \leq \mathbb{E}(Y_{t_1,g_t}(0) | \, D = 1) \leq \mathbb{E}(Y_{t_1,g_t} | \, D = 1)$, this implies that

$$A_4^- \leq A \leq A_4^+,$$

with

$$A_4^- = A_2^-$$

and

$$\begin{aligned} A_4^+ &= \mathbb{E}(Y_{t_1,g_t} | \, D = 1) \times \mathbb{P}(D_{t_1,g_t} = 1) - \mathbb{E}(Y_{t_0,g_t} | \, D = 0) \times (1 - \mathbb{P}(D_{t_0,g_t} = 1)) \\ &- \mathbb{E}(Y_{t_1,g_c} | \, D = 0) \times (1 - \mathbb{P}(D_{t_1,g_c} = 1)) + \mathbb{E}(Y_{t_1,g_t} | \, D = 0) \times (1 - \mathbb{P}(D_{t_1,g_t} = 1)) \\ &+ \mathbb{E}(Y_{t_0,g_c} | \, D = 0) \times (1 - \mathbb{P}(D_{t_0,g_c} = 1)). \end{aligned}$$

Consequently,

$$\max(A_3^-; A_4^-) \leq A \leq \min(A_3^+; A_4^+). \tag{13}$$

Combining (2) and (13) and rearranging yields $B_-^{'}$ and $B_+^{'}$, which are sharp by construction. **QED.**

Proof of Theorem 2.4:
I prove the result for $B_+^{''}$ only.
Let

$$C = ATT_{t_0,g_t} \times \mathbb{P}(D_{t_0,g_t} = 1) + ATT_{t_1,g_c} \times \mathbb{P}(D_{t_1,g_c} = 1) - ATT_{t_0,g_c} \times \mathbb{P}(D_{t_0,g_c} = 1).$$

This is the only quantity appearing in (1) which cannot be estimated from the sample and therefore needs to be bounded to bound ATT_{t_1,g_t}.

If $ATT_{t_1,g_c} = ATT_{t_0,g_c} = ATT_{g_c}$,

$$C = ATT_{t_0,g_t} \times \mathbb{P}(D_{t_0,g_t} = 1) + ATT_{g_c} \left(\mathbb{P}(D_{t_1,g_c} = 1) - \mathbb{P}(D_{t_0,g_c} = 1)\right).$$

Therefore, when $m \leq Y(0) \leq M$,

$$C^- \leq C$$

with

$$\begin{aligned}
&C^- = \mathbb{E}(Y_{t_0,g_t} - M |\, D = 1) \times \mathbb{P}(D_{t_0,g_t} = 1) \\
&+1_{\{t_1\}} \left(\max\left(\mathbb{E}(Y_{t_1,g_c} |\, D = 1); \mathbb{E}(Y_{t_0,g_c} |\, D = 1)\right) - M\right) \left(\mathbb{P}(D_{t_1,g_c} = 1) - \mathbb{P}(D_{t_0,g_c} = 1)\right) \\
&+1_{\{t_0\}} \left(\min\left(\mathbb{E}(Y_{t_1,g_c} |\, D = 1); \mathbb{E}(Y_{t_0,g_c} |\, D = 1)\right) - m\right) \left(\mathbb{P}(D_{t_1,g_c} = 1) - \mathbb{P}(D_{t_0,g_c} = 1)\right).
\end{aligned}$$

A DGP attaining C^- will verify $A = A_5^+$, with

$$\begin{aligned}
&A_5^+ = M \times \mathbb{P}(D_{t_0,g_t} = 1) + \left(1_{\{t_1\}} \min\left(M; M + \Delta\mathbb{E}\right) + 1_{\{t_0\}} \max\left(m; m + \Delta\mathbb{E}\right)\right) \times \mathbb{P}(D_{t_1,g_c} = 1) \\
&+ \left(1_{\{t_1\}} \min\left(M; M - \Delta\mathbb{E}\right) + 1_{\{t_0\}} \max\left(m; m - \Delta\mathbb{E}\right)\right) \times \mathbb{P}(D_{t_0,g_c} = 1).
\end{aligned}$$

Since C is a decreasing function of A,

$$C^- \leq C \Rightarrow A \leq A_5^+$$

But for bounds to be sharp, the DGP attaining them should also verify the common trend assumption, which implies that

$$A \leq A_2^+.$$

Consequently,

$$A \leq \min(A_2^+; A_5^+). \tag{14}$$

Combining (2) and (14) and rearranging yields B_+'', which is sharp by construction.
QED.

Proof of Lemma 2.2:
The proof is the same as the proof of Lemma 2.1, except that all expectations should be taken conditional to X.
QED.

Proof of Theorem 5.1
I start proving equation (11). Let us rewrite ATT_{t_1,g_t}. It is equal to

$$\begin{aligned}
&\int ATT^X_{t_1,g_t} dP(X|D_{t_1,g_t} = 1) \\
= &\int \frac{DID^X_Y + ATT^X_{t_0,g_t} \times \mathbb{P}(D_{t_0,g_t} = 1|X) + ATT^X_{t_1,g_c} \times \mathbb{P}(D_{t_1,g_c} = 1|X) - ATT^X_{t_0,g_c} \times \mathbb{P}(D_{t_0,g_c} = 1|X)}{\mathbb{P}(D_{t_1,g_t} = 1|X)} dP(X|D_{t_1,g_t} = 1)
\end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{\mathbb{P}(D_{t_1,g_t}=1)}\int DID_Y^X + ATT_{t_0,g_t}^X \times \mathbb{P}(D_{t_0,g_t}=1|X) + ATT_{t_1,g_c}^X \times \mathbb{P}(D_{t_1,g_c}=1|X)dP(X_{g_t}) \\
&\quad - \frac{1}{\mathbb{P}(D_{t_1,g_t}=1)}\int ATT_{t_0,g_c}^X \times \mathbb{P}(D_{t_0,g_c}=1|X)dP(X_{g_t}) \\
&= \frac{1}{\mathbb{P}(D_{t_1,g_t}=1)}\int \mathbb{E}(Y_{t_1,g_t}|X) - \mathbb{E}(Y_{t_0,g_t}|X) - [\mathbb{E}(Y_{t_1,g_c}|X) - \mathbb{E}(Y_{t_0,g_c}|X)]\, dP(X_{g_t}) \\
&\quad + \frac{1}{\mathbb{P}(D_{t_1,g_t}=1)}\int ATT_{t_0,g_t}^X \times \mathbb{P}(D_{t_0,g_t}=1|X) + ATT_{t_1,g_c}^X \times \mathbb{P}(D_{t_1,g_c}=1|X) - ATT_{t_0,g_c}^X \times \mathbb{P}(D_{t_0,g_c}=1|X)dP(X_{g_t}) \\
&= \frac{1}{\mathbb{P}(D_{t_1,g_t}=1)}\left[\mathbb{E}(Y_{t_1,g_t}) - \mathbb{E}(Y_{t_0,g_t}) + \mathbb{E}\left((Y_{t_0,g_t}(1) - Y_{t_0,g_t}(0))D\right)\right] - \frac{1}{\mathbb{P}(D_{t_1,g_t}=1)}\int \mathbb{E}(Y_{t_1,g_c}|X) - \mathbb{E}(Y_{t_0,g_c}|X)dP(X_{g_t}) \\
&\quad + \frac{1}{\mathbb{P}(D_{t_1,g_t}=1)}\int ATT_{t_1,g_c}^X \times \mathbb{P}(D_{t_1,g_c}=1|X) - ATT_{t_0,g_c}^X \times \mathbb{P}(D_{t_0,g_c}=1|X)dP(X_{g_t}) \\
&= \frac{1}{\mathbb{P}(D_{t_1,g_t}=1)}\left[\mathbb{E}(Y_{t_1,g_t}) - \mathbb{E}(Y_{t_0,g_t}) - \mathbb{E}(Y_{t_1,g_c}w^X) + \mathbb{E}(Y_{t_0,g_c}w^X)\right] \\
&\quad + \frac{1}{\mathbb{P}(D_{t_1,g_t}=1)}\left[ATT_{t_0,g_t} \times \mathbb{P}(D_{t_0,g_t}=1) + \mathbb{E}\left((Y_{t_1,g_c}(1) - Y_{t_1,g_c}(0))w^X D\right) - \mathbb{E}\left((Y_{t_1,g_c}(1) - Y_{t_1,g_c}(0))w^X D\right)\right] \\
&= \frac{1}{\mathbb{P}(D_{t_1,g_t}=1)}\left[WDID_Y + ATT_{t_0,g_t} \times \mathbb{P}(D_{t_0,g_t}=1) + WATT_{t_1,g_c} \times \mathbb{P}(D_{t_1,g_c}=1) - WATT_{t_0,g_c} \times \mathbb{P}(D_{t_0,g_c}=1)\right].
\end{aligned}
$$

The first equality comes from Lemma 2.2 combined with the fact that the integral is taken over the set of all ω such that $\mathbb{P}(D_{t_1,g_t}=1|X)(\omega) \neq 0$. The second equality is obtained using the definition of conditional probabilities, and using A.2 which implies that $dP(X_{t_1,g_t}) = dP(X_{g_t})$. Finally, the fifth equality is obtained using Bayes' law.

Proof of 1)

In the "no always takers" special case, $\mathbb{P}(D_{t_0,g_t}=1)$, $\mathbb{P}(D_{t_1,g_c}=1)$ and $\mathbb{P}(D_{t_0,g_c}=1)$ are all equal to 0. Therefore, (11) rewrites

$$ATT_{t_1,g_t} = \frac{WDID_Y}{\mathbb{P}(D_{t_1,g_t}=1)},$$

hence the result.

Proof of 2)

If $m \leq Y(0) \leq M$, then

$$\mathbb{E}(Y_{t_0,g_t} - M|D=1) \leq ATT_{t_0,g_t} \leq \mathbb{E}(Y_{t_0,g_t} - m|D=1),$$

$$\mathbb{E}\left((Y_{t_1,g_c} - M)\, w^X|D=1\right) \leq WATT_{t_1,g_c} \leq \mathbb{E}\left((Y_{t_1,g_c} - M)\, w^X|D=1\right)$$

and

$$\mathbb{E}\left((Y_{t_0,g_c} - M)\, w^X|D=1\right) \leq WATT_{t_0,g_c} \leq \mathbb{E}\left((Y_{t_0,g_c} - m)\, w^X|D=1\right).$$

Plugging those three inequalities into (11) yields

$$\frac{WDID_{Y_{-}^{0}}}{\mathbb{P}(D_{t_1,g_t}=1)} \leq ATT_{t_1,g_t} \leq \frac{WDID_{Y_{+}^{0}}}{\mathbb{P}(D_{t_1,g_t}=1)},$$

hence the result.

Proof of 3)

If $m \leq Y(0) \leq Y(1)$, then

$$0 \leq ATT_{t_0,g_t} \leq \mathbb{E}(Y_{t_0,g_t} - m|D=1),$$

$$0 \leq WATT_{t_1,g_c} \leq \mathbb{E}\left((Y_{t_1,g_c} - M)\, w^X | D=1\right)$$

and

$$0 \leq WATT_{t_0,g_c} \leq \mathbb{E}\left((Y_{t_0,g_c} - m)\, w^X | D=1\right).$$

Plugging those three inequalities into (11) yields

$$\frac{WDID_{Y_{-}^{1}}}{\mathbb{P}(D_{t_1,g_t}=1)} \leq ATT_{t_1,g_t} \leq \frac{WDID_{Y_{+}^{1}}}{\mathbb{P}(D_{t_1,g_t}=1)},$$

hence the result.

QED.

Proof of Theorem 5.1

I only show how to derive $WB_{-}^{''}$. Let

$$\min(X) = \min\left(m_1(X); m_2(X)\right)$$

where

$$m_1(X) = \mathbb{E}(Y_{t_1,g_c} - M|D=1,X)\left(\mathbb{P}(D_{t_1,g_c}=1|X) - \mathbb{P}(D_{t_0,g_c}=1|X)\right)$$

and

$$m_2(X) = \mathbb{E}(Y_{t_1,g_c} - m|D=1,X)\left(\mathbb{P}(D_{t_1,g_c}=1|X) - \mathbb{P}(D_{t_0,g_c}=1|X)\right).$$

Let

$$\mathbb{E}_{\min} = \mathbb{E}\left(\min\left((Y_{t_1,g_c} - M)\left(1-p^X\right); (Y_{t_1,g_c} - m)\left(1-p^X\right)\right) w^X | D=1\right).$$

Finally, let $1_{\{t_1\}}(X)$ and $1_{\{t_0\}}(X)$ be the indicators of the events $\mathbb{P}(D_{t_1,g_c}=1|X) > \mathbb{P}(D_{t_0,g_c}=1|X)$ and $\mathbb{P}(D_{t_1,g_c}=1|X) < \mathbb{P}(D_{t_0,g_c}=1|X)$.

If $ATT^X_{t_1,g_c} = ATT^X_{t_0,g_c}$, (6) rewrites as

$$ATT^X_{t_1,g_t} = \frac{DID^X_Y + ATT^X_{t_0,g_t} \times \mathbb{P}(D_{t_0,g_t}=1|X) + ATT^X_{t_1,g_c}\left(\mathbb{P}(D_{t_1,g_c}=1|X) - \mathbb{P}(D_{t_0,g_c}=1|X)\right)}{\mathbb{P}(D_{t_1,g_t}=1|X)}.$$

Therefore,

$$\frac{DID^X_Y + \mathbb{E}(Y_{t_0,g_t} - M|D=1,X) \times \mathbb{P}(D_{t_0,g_t}=1|X) + \min(X)}{\mathbb{P}(D_{t_1,g_t}=1|X)}$$

is a lower bound to $ATT^X_{t_1,g_c}$. Plugging this lower bound into (5) yields that

$$\int \frac{DID^X_Y + \mathbb{E}(Y_{t_0,g_t} - M|D=1,X) \times \mathbb{P}(D_{t_0,g_t}=1|X) + \min(X)}{\mathbb{P}(D_{t_1,g_t}=1|X)} dP(X|D_{t_1,g_t}=1)$$

is a lower bound for ATT_{t_1,g_t}. Using the same steps as those used to prove equation (11), one can show that this lower bound rewrites

$$\frac{WDID_Y + ATT_{t_0,g_t} \times \mathbb{P}(D_{t_0,g_t}=1) + \int \min(X) dP(X_{g_t})}{\mathbb{P}(D_{t_1,g_t}=1)}. \tag{15}$$

Then,

$$\begin{aligned}
& \int \min(X) dP(X_{g_t}) \\
= & \int 1_{\{t_1\}}(X) \mathbb{E}(Y_{t_1,g_c} - M|D=1,X) \times \left(\mathbb{P}(D_{t_1,g_c}=1|X) - \mathbb{P}(D_{t_0,g_c}=1|X)\right) dP(X_{g_t}) \\
& + \int 1_{\{t_0\}}(X) \mathbb{E}(Y_{t_1,g_c} - m|D=1,X) \times \left(\mathbb{P}(D_{t_1,g_c}=1|X) - \mathbb{P}(D_{t_0,g_c}=1|X)\right) dP(X_{g_t}) \\
= & \int \mathbb{E}\left(1_{\{t_1\}}(X)\left(Y_{t_1,g_c} - M\right)\left(1-p^X\right)|D=1,X\right) \mathbb{P}(D_{t_1,g_c}=1|X) dP(X_{g_t}) \\
& + \int \mathbb{E}\left(1_{\{t_0\}}(X)\left(Y_{t_1,g_c} - m\right)\left(1-p^X\right)|D=1,X\right) \mathbb{P}(D_{t_1,g_c}=1|X) dP(X_{g_t}) \\
= & \int \mathbb{E}\left(1_{\{t_1\}}(X)\left(Y_{t_1,g_c} - M\right)\left(1-p^X\right) D|X\right) dP(X_{g_t}) \\
& + \int \mathbb{E}\left(1_{\{t_0\}}(X)\left(Y_{t_1,g_c} - m\right)\left(1-p^X\right) D|X\right) dP(X_{g_t}) \\
= & \int \mathbb{E}\left(1_{\{t_1\}}(X)\left(Y_{t_1,g_c} - M\right)\left(1-p^X\right) Dw^X|X\right) dP(X_{g_c}) \\
& + \int \mathbb{E}\left(1_{\{t_0\}}(X)\left(Y_{t_1,g_c} - m\right)\left(1-p^X\right) Dw^X|X\right) dP(X_{g_c}) \\
= & \mathbb{E}\left(1_{\{t_1\}}(X)\left(Y_{t_1,g_c} - M\right)\left(1-p^X\right) Dw^X\right) + \mathbb{E}\left(1_{\{t_0\}}(X)\left(Y_{t_1,g_c} - m\right)\left(1-p^X\right) Dw^X\right) \\
= & \mathbb{E}\left(\min\left(\left(Y_{t_1,g_c} - M\right)\left(1-p^X\right); \left(Y_{t_1,g_c} - m\right)\left(1-p^X\right)\right) w^X|D=1\right) \times \mathbb{P}(D_{t_1,g_c}=1). \\
= & \mathbb{E}_{\min} \times \mathbb{P}(D_{t_1,g_c}=1).
\end{aligned}$$

The first equality comes from the fact that

$$\begin{aligned}\min(X) = 1_{\{t_1\}}(X)\mathbb{E}(Y_{t_1,g_c} - M|D=1,X)\left(\mathbb{P}(D_{t_1,g_c}=1|X) - \mathbb{P}(D_{t_0,g_c}=1|X)\right)\\ +1_{\{t_0\}}(X)\mathbb{E}(Y_{t_1,g_c} - m|D=1,X) \times \left(\mathbb{P}(D_{t_1,g_c}=1|X) - \mathbb{P}(D_{t_0,g_c}=1|X)\right)\end{aligned}$$

because $\min(X) = 0$ when $\mathbb{P}(D_{t_1,g_c}=1|X) = \mathbb{P}(D_{t_0,g_c}=1|X)$. The second equality hold because I have assumed that $\mathbb{P}(D_{t_1,g_c}=1|X) > 0 \Leftrightarrow \mathbb{P}(D_{t_0,g_c}=1|X) > 0$ with probability one. Finally, the fourth equality holds because of Bayes' law. Combining the last equality with (15), I get that

$$\frac{WDID_Y + ATT_{t_0,g_t} \times \mathbb{P}(D_{t_0,g_t}=1) + \mathbb{E}_{\min} \times \mathbb{P}(D_{t_1,g_c}=1)}{\mathbb{P}(D_{t_1,g_t}=1)}$$

is a lower bound for ATT_{t_1,g_t}. This lower bound can be rewritten:

$$\frac{WDID_{Y_-^3}}{\mathbb{P}(D_{t_1,g_t}=1)}.$$

If $ATT^X_{t_1,g_c} = ATT^X_{t_0,g_c}$, (6) also rewrites as

$$ATT^X_{t_1,g_t} = \frac{DID^X_Y + ATT^X_{t_0,g_t} \times \mathbb{P}(D_{t_0,g_t}=1|X) + ATT^X_{t_0,g_c}\left(\mathbb{P}(D_{t_1,g_c}=1|X) - \mathbb{P}(D_{t_0,g_c}=1|X)\right)}{\mathbb{P}(D_{t_1,g_t}=1|X)}.$$

Then, following the same steps as above, one can show that

$$\frac{WDID_{Y_-^4}}{\mathbb{P}(D_{t_1,g_t}=1)}$$

is also a lower bound for ATT_{t_1,g_t}, hence the result.

QED.

Proof of Theorem 3.1

The moment inequality model in Andrews & Soares (2010) is based on the following assumptions. The parameter of interest θ_0 and the true distribution of the data F_0 are assumed to belong to a parameter space $\mathcal{F} = (\theta, F)$ such that:

i) $\theta \in \mathbb{R}$

ii) The data W_i are iid.

iii) $\mathbb{E}_F\left(m_j\left(W,\theta,\eta\right)\right) \geq 0$ for $j = 1,...,p$, where m_j are known real-valued functions, and η is a parameter. η should be identified, and there should exist a consistent and asymptotically normal estimator of it.

iv) $\sigma^2_{F,j}\left(\theta\right) = \mathbb{V}_F\left(m_j\left(W,\theta,\eta\right)\right) \in (0,+\infty)$

Let $m\left(W,\theta,\eta\right) = \left(m_1\left(W,\theta,\eta\right),...,m_p\left(W,\theta,\eta\right)\right)'$,

$\Sigma_F\left(\theta\right) = \mathbb{V}_F\left(m\left(W,\theta,\eta\right)\right)$,

$D_F(\theta) = Diag(\Sigma_F)$,

$\Omega_F(\theta) = D^{-\frac{1}{2}}(\theta)\,\Sigma(\theta)\,D^{-\frac{1}{2}}(\theta)$,

$\overline{m}_{n,j}(\theta) = \frac{1}{n}\sum_{i=1}^{n} m_j(W_i, \theta, \widehat{\eta}_n(\theta))$ for $j = 1, ..., p$,

$\overline{m}_n(\theta) = (\overline{m}_{n,1}(\theta), ..., \overline{m}_{n,p}(\theta))'$,

$\widehat{\Sigma}_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} (m(W_i, \theta, \widehat{\eta}_n) - \overline{m}_n(\theta))(m(W_i, \theta, \widehat{\eta}_n) - \overline{m}_n(\theta))'$,

$\widehat{\sigma}_{n,j}(\theta) = \left[\widehat{\Sigma}_n(\theta)\right]_{j,j}$,

and

$\widehat{D}_n(\theta) = Diag\left(\widehat{\Sigma}_n(\theta)\right)$.

For all sequences (θ_n, F_n) in $\mathcal{F}$ such that

$\theta_n \underset{n\to+\infty}{\to} \theta$,

$F_n \underset{n\to+\infty}{\to} F$,

$\Omega_{F_n} \underset{n\to+\infty}{\to} \Omega_F$,

and

$\sqrt{n}\sigma^{-1}_{F_{n,j}}(\theta_n)\,\mathbb{E}_{F_n}(m_j(W, \theta_n, \eta)) \underset{n\to+\infty}{\to} h_j \in \overline{\mathbb{R}}_+$, we should have that

v) $A_n = (A_{n,1}, ..., A_{n,p})' \underset{d}{\to} Z \sim \mathcal{N}(0_k, \Omega_F)$ as $n \to +\infty$, where

$$A_{n,j} = \sqrt{n}\sigma^{-1}_{F_{n,j}}\left(\overline{m}_{n,j}(\theta_n) - \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{F_n}(m_j(W, \theta_n, \eta))\right)$$

vi) $\frac{\widehat{\sigma}_{n,j}(\theta_n)}{\sigma_{F_n,j}(\theta_n)} \underset{p}{\to} 1$ when $n \to +\infty$ for $j = 1$ to p.

vii) $\widehat{D}_n^{-\frac{1}{2}}(\theta_n)\,\widehat{\Sigma}_n(\theta_n)\,\widehat{D}_n^{-\frac{1}{2}}(\theta_n) \to \Omega_F$ when $n \to +\infty$.

As Andrews & Guggenberger (2010) show in their 2nd lemma, a sufficient condition for v), vi) and vii) to hold is the following condition:

viii) $\mathbb{E}_F\left(|m_j(W, \theta, \eta)|^{2+\delta}\right) < K$ for some $\delta > 0$, where K is a constant.

Finally, for the confidence interval not to be conservative, the following assumption should be verified:

ix) $\mathbb{E}_F(m_j(W, \theta, \eta)) = 0$ for some j and some $(\theta, F) \in \mathcal{F}$.

Therefore, I show now that my model can be rewritten as a moment inequality model which verifies assumptions i)-iv) and viii)-ix). $W = (Y, T, G, D)$. Let $\eta_1 = \mathbb{P}(T = 1, G = 1)$, $\eta_2 = \mathbb{P}(T = 0, G = 1)$, $\eta_3 = \mathbb{P}(T = 1, G = 0)$, $\eta_4 = \mathbb{P}(T = 0, G = 0)$ and $\eta = (\eta_1, \eta_2, \eta_3, \eta_4)$. As shown in section 2, Y^0_- and Y^0_+ can be written as functions f^0_- and f^0_+ of the data. Let $g(x, \eta) = \frac{TGx}{\eta_1} - \frac{(1-T)Gx}{\eta_2} - \frac{T(1-G)x}{\eta_3} + \frac{(1-T)(1-G)x}{\eta_4}$.

Assumption i) and ii) hold because $ATT_{t_1,g_t} \in \mathbb{R}$ and the data is iid.

(10) is equivalent to

$$\mathbb{E}_{F_0}\left(m_j\left(W, \theta_0, \eta\right)\right) \geq 0$$

for $j = 1$ to 4, with

$$\begin{aligned}
m_1\left(W, \theta_0, \eta\right) &= \tfrac{TGD}{\eta_1}\theta_0 - g(f_-^0(W), \eta) \\
m_2\left(W, \theta_0, \eta\right) &= TGD\left(\theta_0 - (Y - M)\right) \\
m_3\left(W, \theta_0, \eta\right) &= g(f_+^0(W), \eta) - TGD\theta_0 \\
m_4\left(W, \theta_0, \eta\right) &= TGD\left(Y - m - \theta_0\right).
\end{aligned}$$

Moreover, η is identified and can be consistently estimated with an asymptotically normal estimator. This implies that iii) holds.

Using the triangle inequality, the fact that $Y(0)$ and $Y(1)$ are bounded, that $\min(\eta_1, \eta_2, \eta_3, \eta_4) \geq \epsilon$ and that $x \mapsto x^{2+\delta}$ is increasing on $\mathbb{R}^+$, it is easy to show that $\mathbb{E}_F\left(\left|m_j\left(W, \theta, \eta\right)\right|^{2+\delta}\right) \leq K^j$ where the K^j write as functions of m, M, m', M', δ and ϵ. Setting $K = \max\left(K_1, K_2, K_3, K_4\right) + 1$, viii) is verified.

viii) implies that $\mathbb{E}_F\left(\left|m_j\left(W, \theta, \eta\right)\right|^2\right) < +\infty$ so that $\sigma_{F,j}^2\left(\theta\right) < +\infty$, and $0 < \sigma_{F,j}^2\left(\theta\right)$ for non degenerate distributions. Therefore iv) holds.

Finally, sharpness of the bounds implies that ix) holds.

QED.