Centre de Recherche en Économie et Statistique
CREST

Série des Documents de Travail

n° 2011-10

Fuzzy Differences in Differences

C. DE CHAISEMARTIN[1]

[1] Paris School of Economics, 48 boulevard Jourdan, 75014 Paris France. chaisemartin@pse.ens.fr
et CREST, 15 boulevard Gabriel Péri, 92245 Malakoff, France.

Fuzzy Differences in Differences[*]

Clément de Chaisemartin[†][‡]

March 22, 2011

Abstract

Difference in differences (DID) require that the treatment rate is 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"). Sometimes, treatment rate increases more in the treatment than in the control group but there are never or always takers. This paper derives identification results applying to such settings. They only require one common trend assumption on the outcome of interest (Y) whereas the standard instrumental variable result usually invoked also requires common trend on treatment rate. When there are never takers but no or few always takers, common trend on Y is sufficient to identify exactly an ATT or at least its sign.

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

JEL Codes: C21, C23, I19

[*] I am extremely thankful to Xavier d'Haultfoeuille for his numerous advice and encouragements. I am very grateful to Luc Behaghel, Andrew Clark, Joseph Doyle, Pauline Givord, Marc Gurgand, Edwin Leuven, Thierry Magnac, Eric Maurin, Thomas Piketty, Roland Rathelot, participants at the European Conference in Econometrics (EC^2), at the French Econometrics Conference, seminar participants at the Paris School of Economics and at CREST for their helpful comments.

[†] Paris School of Economics, 48 boulevard Jourdan 75014 Paris France, chaisemartin@pse.ens.fr

[‡] CREST, 15 boulevard Gabriel Péri 92245 Malakoff France

1 Introduction

Since the seminal work by Ashenfelter and Card [1985], 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 some mean outcome Y between two periods (0 and 1) and across two groups of individuals (control and treatment). In Rubin's causal model where potential outcomes with and without treatment ($Y(1)$ and $Y(0)$) are introduced, and where treatment effects ($Y(1) - Y(0)$) are allowed to be heterogeneous across observations, it has been shown that a DID identifies an ATT under two assumptions (see Abadie [2005]). 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. 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 (or treatment intensity if treatment is multivariate) increases more in the treatment than in the control group but there are "never" or "always" takers.[2] This differential change in treatment rate / intensity across the control and the treatment group might still be used to identify an ATT. This is what I refer to as a fuzzy DID identification strategy.

When compliance is imperfect, common trend alone is not sufficient for identification in a model allowing for heterogeneous treatment effect. Under common trend on $Y(0)$, if no observation is treated in any group, trends are parallel in the two groups and the DID is merely equal to 0. In a standard DID, the only reason why trends might diverge across groups is that observations in the treatment group × period 1 cell get treated, so that the DID measures the effect of the treatment on them. A

[1] By never takers, I merely refer to untreated observations in the treatment group in period 1. Always takers are treated observations in the three other groups.

[2] This might not be an issue when panel data is available. In this case, researchers can indeed *choose* observations making up the treatment and the control group. They can for instance keep only observations of the control group who were untreated in period 0 and 1, and observations of the treatment group untreated in period 0 and treated in period 1 (see for instance Field [2005]). Despite the arbitrariness of this definition of groups, which might entail that common trend is violated, it ensures that the perfect compliance assumption is met. But when only pooled cross-sections are available, it is no longer possible to select observations thus. Even in the panel data case, such a trick might not be possible when treatment is multivariate and that all observations received some amount of treatment. Think of treatment as number of years of education completed: hardly anyone completes 0 years of education.

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 × 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 effect of the treatment might vary across cells. Assuming $Y(1) - Y(0)$ to be constant across observations would solve the issue: the unknowns in the DID equation would all be equal to each other so that we would be back to one equation with one unknown.

Therefore, the starting point of the paper is to show that in a fuzzy DID, when treatment effect is allowed to be heterogeneous, a common trend assumption on $Y(0)$ is generally not sufficient to identify some ATT. However, in the special case where there are never takers but no always takers (a situation I henceforth refer to as the "no always takers" special case), this assumption is sufficient for identification, as in the standard DID model. Indeed, in such a situation, even though not all observations of the test group are treated in period 1, there are still treated observations in one group only, so that a DID computation will yield one equation with one unknown. When there are always takers, common trend on $Y(0)$ does not allow for point identification, but partial identification of some ATT is still possible provided Y is bounded. I derive explicit sharp bounds in this case. The identification region is likely to be narrow enough to identify the sign of this ATT when there are "few" always takers. Whether there are "many" or "few" never takers does not matter. It is also possible to derive a second and narrower identification region for the same ATT under the supplementary assumption that treatment effects do not change between the two periods in the control group. This second identification region will be narrow when there are few treated observations in the treatment group in period 0, and when the change in the the treatment rate from period 0 to 1 is small in the control group.

Actually, fuzzy DID has already been used often in the applied economics literature. Up to now, researchers who implemented it 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. Duflo [2001] uses this strategy to estimate the impact of educational attainment on wages. Bleakley and Chin [2004] use it to estimate the impact of language proficiency on wages. Papers which use differential evolution of exposure to treatment across US states to estimate treatment effects build upon the same intuition. A good example is Evans and 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 children's weight. Because their regressions include state and year fixed effects, their estimate arises from the comparison of the evolution of children's weight in states with and without changes in tax. Had there been only two states and two years in their data, their regression coefficient would merely be the DID on children's weight divided by the DID on smoking prevalence among pregnant women, as in Duflo [2001] or Bleakley and Chin [2004]. However, the underlying assumptions of this identification strategy have not been clarified so far.

Imbens and Angrist [1994] have shown that IV coefficients can be interpreted as a local average treatment effect (LATE) in a model allowing for heterogeneous treatment effect. I put forward in a companion paper (de Chaisemartin [2010]) that when applied to fuzzy DID their result holds under two common trend assumptions, on Y and on D, and a monotonicity assumption (no "defiers"). Common trend on Y allows recovering the intention to treat effect of the policy, whereas common trend on D allows recovering the share of compliers.

My results contribute to the literature because they require only one common trend assumption on Y. Thus, I remove the monotonicity condition. Even though it is often thought of as an innocuous assumption, it may be restrictive in some instances as discussed in Small and Tan [2007]. Above all, they do not require common trend on D. One might argue that the marginal cost of this second common trend assumption is weaker than for the first: if one is ready to believe that without the program trends would have been parallel on Y, one should be ready to take the same assumption on D. However, this might not always be true. For instance, in Evans and Ringel [1999], 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 babies weight at birth. Moreover, even in applications where there is no obvious reason to suspect that trends on Y or on D would have strongly diverged, there is no reason why they should have been **exactly** parallels neither because assignment to treatment is not random. The most we can reasonably expect is that trends in the untreated group provide a fairly good "first order approximation" of what would have happened in the treated group. Results requiring one first order approximation might therefore be more reliable than results requiring two. The combination of two small errors in the numerator and in the denominator of the Wald-DID could indeed lead to a large difference between the Wald-DID and the true treatment effect. Therefore, the contribution of this paper is to bring new fuzzy DID identification results which rely on weaker assumptions than the standard Imbens and Angrist IV result.

My results might be useful in applications with no or few always takers. To illustrate this, I measure the efficacy of a new pharmacotherapy for smoking cessation. Varenicline is a drug which was made available to French cessation clinics in February 2007 as one possible pharmacotherapy for smoking cessation support. In 15 services, less than 3% of all patients consulted have been prescribed varenicline during the year following its release. In 13 services, more than 20% of patients were prescribed varenicline. Because in this application there are some but few always takers, I derive bounds for the ATT which are narrow enough to infer its sign. Had there been more always takers, 0 would lie within the identification region. Therefore, in a fuzzy DID, common trend on Y is sufficient to obtain accurate information on an ATT when there are few always takers, even if there are many never takers. My results might also be useful in applications considering the extension of a policy, that is to say when the control group was already eligible in period 0 and the test group became eligible in period 1 (see for instance Bach [2009]). Indeed, in such situations the share of treated observations in the treatment group in period 1 is by definition equal to 0. Consequently, the second identification region I derive will be narrow provided the share of treated observations did not change too much between period 0 and 1 in the control group.

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

2 Identification

I place myself in the pooled cross-section case: each individual is observed only at one period. 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. I assume that treatment status is binary and is denoted by an indicator D (I show in Appendix A that all results can easily be extended when treatment takes a finite number of values).

Throughout the paper it is implicitly assumed that the stable unit treatment value assumption holds. Under this assumption I define $Y(1)$ and $Y(0)$ as 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 treatment effect is $Y(1) - Y(0)$. Average treatment effects are the corresponding expectations. $X \sim Y$ means that X and Y have the same probability distribution. $\mathbb{X}$ is the support of X. To alleviate the notational burden, I introduce several shorthands following Athey and Imbens [2006]:

$Y_{ij}(k) \sim Y(k) \,|\, t = i,\, g = j \;\forall (k, i, j) \in \{0; 1\} \times \{t_0; t_1\} \times \{g_c; g_t\}$

$Y_{ij} \sim Y \,|\, t = i,\, g = j \;\forall (i, j) \in \{t_0; t_1\} \times \{g_c; g_t\}$

$D_{ij} \sim D \,|\, t = i,\, g = j \;\forall (i, j) \in \{t_0; t_1\} \times \{g_c; g_t\}$

Under those notations, the standard DID parameter is:

$$DID = \mathbb{E}(Y_{t_1,g_t}) - \mathbb{E}(Y_{t_0,g_t}) - \left[\mathbb{E}(Y_{t_1,g_t}) - \mathbb{E}(Y_{t_0,g_c})\right].$$

I denote by DID^P the DID on treatment rate from period 0 to 1 across the two groups. I assume that $DID^P \neq 0$: the definition of a fuzzy DID is that exposure to treatment should have evolved differentially in the two groups. Without loss of generality, I assume that $DID^P > 0$. The no always takers special case is met when $\mathbb{P}(D_{t_0,g_t} = 1) = \mathbb{P}(D_{t_1,g_c} = 1) = \mathbb{P}(D_{t_0,g_c} = 1) = 0$. It is likely to arise for instance when a new social program is implemented with only a specific group eligible to it (unemployed...) and take-up is below 100%. $ATT_{i,j} = \mathbb{E}(Y_{i,j}(1) - Y_{i,j}(0) |\, D = 1), \forall (i, j) \in \{t_0; t_1\} \times \{g_c; g_t\}$ is the average treatment effect on treated individuals of group j in period i. $ATT = \mathbb{E}(Y(1) - Y(0) |\, D = 1)$ is the average treatment effect on the treated. I denote $\mathbb{P}_{AT} = \mathbb{P}(D_{t_0,g_t} = 1) + \mathbb{P}(D_{t_1,g_c} = 1) + \mathbb{P}(D_{t_0,g_c} = 1)$ the sum of the three shares of always takers.

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

A.1 Common trend for the outcome variable

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

Lemma 1: Non-identification.

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

$$DID = ATT_{t_1,g_t} \times \mathbb{P}(D_{t_1,g_t} = 1) - ATT_{t_0,g_t} \times \mathbb{P}(D_{t_0,g_t} = 1)$$

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

According to Lemma 1, under A.1, if compliance is imperfect, the DID on Y can be written as a weighted DID of four average treatment effects on four different populations. This is the equation with several unknowns mentioned in the introduction. Because two ATT enter the equation with positive sign and two enter with negative sign, the DID cannot be given any causal interpretation. It might for instance be positive whereas the four ATT are negative. The intuition for this result is that under common trend on $Y(0)$, if no observations had been treated in any of the four time × group cells, trends would have been parallel in the two groups and the DID would have merely been equal to 0. In a standard DID, the only reason why trends might diverge across groups is that observations in the treatment group get treated in period 1, so that the DID measures the effect of the treatment on them. In a fuzzy DID, since there might be treated observations in several time × 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 subgroups, it is not possible to identify any of the $ATT_{i,j}$ from a standard DID computation, since it yields one equation with several unknowns.

Proposition 1: Point identification.

i) In the no always takers special case, A.1 is sufficient for ATT_{t_1,g_t} to be identified and

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

ii) Under A.1 and the supplementary assumption that $\forall (i,j) \in \{t_0;t_1\} \times \{g_c;g_t\}$, $ATT_{i,j} = ATT$, the $ATT_{i,j}$ and the ATT are identified:

$$\forall (i,j) \in \{t_0;t_1\} \times \{g_c;g_t\}, \; ATT_{i,j} = ATT = \frac{DID}{DID^P}$$

In the no always takers special case, common trend is sufficient to identify ATT_{t_1,g_t} as in a standard DID because there are treated observations in one group only. Therefore, there is only one unknown left in (1). This result is strikingly similar to Battistin and Rettore's [2008] result on regression discontinuity (RDD). 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. Estimation of ATT_{t_1,g_t} still requires being able to estimate $\mathbb{P}(D_{t_1,g_t} = 1)$. Sometimes treatment status is not observed, making it impossible to estimate $\mathbb{P}(D_{t_1,g_t} = 1)$ (see e.g. Eissa and Leibman [1996]). Since ATT_{t_1,g_t} and DID have the same sign and $|DID| \leq |ATT_{t_1,g_t}|$, it is at least possible to estimate a lower bound of ATT_{t_1,g_t} by computing the DID. For instance, Eissa and Leibman's 1.4 percentage points DID is a lower bound on the true effect of the EITC extension on lone mothers' participation to the labor market.

Then in part ii) of Proposition 1, I show that it is enough to restrict the heterogeneity of the treatment effect, assuming that it does not vary across time and group, to identify exactly the ATT. This is because under this assumption the four unknowns in (1) are actually 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 the case 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 groups.

Before stating Proposition 2, I define three quantities:

$B^0(u,v) = \frac{DID+\big(\mathbb{E}(Y_{t_0,g_t}|\,D=1)-u\big)\times\mathbb{P}(D_{t_0,g_t}=1)+\big(\mathbb{E}(Y_{t_1,g_c}|\,D=1)-u\big)\times\mathbb{P}(D_{t_1,g_c}=1)-\big(\mathbb{E}(Y_{t_0,g_c}|\,D=1)-v\big)\times\mathbb{P}(D_{t_0,g_c}=1)}{\mathbb{P}(D_{t_1,g_t}=1)}$,

$B^1 = \frac{DID+\big(\mathbb{E}(Y_{t_0,g_t}|\,D=1)-M\big)\times\mathbb{P}(D_{t_0,g_t}=1)+\big(max\big(\mathbb{E}(Y_{t_1,g_c}|\,D=1);\mathbb{E}(Y_{t_0,g_c}|\,D=1)\big)-M\big)\times\big(\mathbb{P}(D_{t_1,g_c}=1)-\mathbb{P}(D_{t_0,g_c}=1)\big)}{\mathbb{P}(D_{t_1,g_t}=1)}$ and

$B^2 = \frac{DID+\big(\mathbb{E}(Y_{t_0,g_t}|\,D=1)-m\big)\times\mathbb{P}(D_{t_0,g_t}=1)+\big(min\big(\mathbb{E}(Y_{t_1,g_c}|\,D=1);\mathbb{E}(Y_{t_0,g_c}|\,D=1)\big)-m\big)\times\big(\mathbb{P}(D_{t_1,g_c}=1)-\mathbb{P}(D_{t_0,g_c}=1)\big)}{\mathbb{P}(D_{t_1,g_t}=1)}$.

Proposition 2: Partial Identification.

i) Under A.1 and the supplementary assumption that $\exists(m,\,M)\in\mathbb{R}^2/\,\mathbb{P}(m\leq Y(0)\leq M)=1$,

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

$B_-=max\,\big(B^0(M,m)\;;\;\mathbb{E}(Y_{t_1,g_t}|\,D=1)-M\big)$ *and* $B_+=min\,\big(B^0(m,M)\;;\;\mathbb{E}(Y_{t_1,g_t}|\,D=1)-m\big)$,

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

$\mathbb{P}_{AT}\leq\mathbb{P}(D_{t_1,g_t}=1)$ *is a sufficient condition to have that either* $B_-=B^0(M,m)$ *or* $B_+=B^0(m,M)$.

ii) Under A.1 and the supplementary assumptions that $\exists(m,\,M)\in\mathbb{R}^2/\,\mathbb{P}(m\leq Y(0)\leq M)=1$ *and that* $ATT_{t_1,g_c}=ATT_{t_0,g_c}$,

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

$B_-^{'}=max\,\big(min(B^1\;;\;B^2)\;;\;\mathbb{E}(Y_{t_1,g_t}|\,D=1)-M\big)$ *and*

$B_+^{'}=min\,\big(max(B^1\;;\;B^2)\;;\;\mathbb{E}(Y_{t_1,g_t}|\,D=1)-m\big)$.

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

$\mathbb{P}(D_{t_0,g_t}=1)+|\mathbb{P}(D_{t_1,g_c}=1)-\mathbb{P}(D_{t_0,g_c}=1)|\leq\mathbb{P}(D_{t_1,g_t}=1)$ *is a sufficient condition to have that either* $B_-^{'}=min(B^1\;;\;B^2)$ *or* $B_+^{'}=max(B^1\;;\;B^2)$.

If $Y(0)$ is bounded, it is possible to find bounds for ATT_{t_1,g_t} which can be non-parametrically estimated from the sample in the spirit of Manski [1990]. This comes from the fact that the only three quantities appearing in (1) which are not observed and do not belong to ATT_{t_1,g_t} 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, it suffices to build up worst-case scenarii for each of them to derive bounds for ATT_{t_1,g_t}. But those worst case scenarii might not be compatible with the common trend assumption and might therefore yield values lower (resp. higher) than the lowest (resp. highest) possible value for ATT_{t_1,g_t} compatible with the data, i.e. $\mathbb{E}(Y_{t_1,g_t}|\, D = 1) - M$ (resp. $\mathbb{E}(Y_{t_1,g_t}|\, D = 1) - m$). Hence the need to ensure that $B_- \geq \mathbb{E}(Y_{t_1,g_t}|\, D = 1) - M$ and $B_+ \leq \mathbb{E}(Y_{t_1,g_t}|\, D = 1) - m$. If $B_- = \mathbb{E}(Y_{t_1,g_t}|\, D = 1) - M$ and $B_+ = \mathbb{E}(Y_{t_1,g_t}|\, D = 1) - m$, the bounds are uninformative. 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 period 1 × treatment group cell is greater than the shares of always takers, 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 in uninformative. There is no sufficient condition on $\mathbb{P}_{AT}$ which ensures that the two bounds are informative (except $\mathbb{P}_{AT} = 0$), because even when $\mathbb{P}_{AT}$ is very small, it is still possible to build up a DGP such that one of the bounds is uninformative, for instance setting $\mathbb{E}(Y_{t_0,g_t}(0)|\, D = 1) = M$. Apart from such extreme cases, if $\mathbb{P}_{AT} \leq \mathbb{P}(D_{t_1,g_t} = 1)$, it is likely that the two bounds will be informative. This condition appears because $\mathbb{P}_{AT}$ is the "size" of the three subgroups for which $Y(0)$ is not observed, which enter into (1), and for which worst case scenarii must be constructed. $\mathbb{P}(D_{t_1,g_t} = 1)$ is the size of the only subgroup for which $Y(0)$ is not observed, which enters the common trend equation and does not enter into (1), that is to say the size of the only degree of freedom left to verify common trend once worst case scenarii have been constructed for the three groups of always takers. When the two bounds are informative, the length of $[B_-; B_+]$ is equal to $(M - m) \times \frac{\mathbb{P}_{AT}}{\mathbb{P}(D_{t_1,g_t}=1)}$. It is increasing with $\mathbb{P}_{AT}$, and decreasing with $\mathbb{P}(D_{t_1,g_t} = 1)$. However, whether 0 belongs to $[B_-; B_+]$ does not depend on $\mathbb{P}(D_{t_1,g_t} = 1)$ but on the size of DID with respect to $M - m$, , $\mathbb{P}(D_{t_0,g_t} = 1)$, $\mathbb{P}(D_{t_1,g_c} = 1)$, and $\mathbb{P}(D_{t_0,g_c} = 1)$.

In part ii) of Proposition 2 I show that narrower bounds for ATT_{t_1,g_t} can be derived under the supplementary assumption that the ATT is constant over time in the control group.[3] Such an assumption might be credible for instance when the treatment rate does not significantly change between period 0 and 1 in the control group, when observable characteristics of treated individuals in the control group do not change much over the two periods, or when $\mathbb{E}(Y_{t_1,g_c}|\, D = 1)$ is close from $\mathbb{E}(Y_{t_0,g_c}|\, D = 1)$. Under this hypothesis, (1) becomes an equation with only three unknowns, and worst case analysis must be conducted on only two expectations. Those worst case scenarii might also not be compatible with common trend and may therefore yield lower and upper bounds

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

outside the range of values of ATT_{t_1,g_t} compatible with the data, hence the need to ensure that $B^{'}_{-} \geq \mathbb{E}(Y_{t_1,g_t} | D = 1) - M$ and $B^{'}_{+} \leq \mathbb{E}(Y_{t_1,g_t} | D = 1) - m$ for the bounds to be sharp. If $\mathbb{P}(D_{t_0,g_t} = 1) + |\mathbb{P}(D_{t_1,g_c} = 1) - \mathbb{P}(D_{t_0,g_c} = 1)| \leq \mathbb{P}(D_{t_1,g_t} = 1)$, at least one of the two bounds will be informative. The sign of ATT_{t_1,g_t} will be identified if both $\mathbb{P}(D_{t_0,g_t} = 1)$ and $|\mathbb{P}(D_{t_1,g_c} = 1) - \mathbb{P}(D_{t_0,g_c} = 1)|$ are small. With respect to part i) of the Proposition, $\mathbb{P}(D_{t_1,g_c} = 1) + \mathbb{P}(D_{t_0,g_c} = 1)$ has been replaced by $|\mathbb{P}(D_{t_1,g_c} = 1) - \mathbb{P}(D_{t_0,g_c} = 1)|$: what matter are no longer the shares of treated observations in the control group but the change in this share from period 0 to 1. This is somewhat similar to the change in the size of the identification region when Lee bounds (see Lee [2009] and Horowitz and Manski [1995]) are used to deal with attrition instead of Manski bounds. This result is of particular interest to place narrow bounds on the ATT in applications considering the extension of policy to a group which was previously not eligible to it and which use a group previously eligible as the control group. Indeed, in such cases $\mathbb{P}(D_{t_0,g_t} = 1) = 0$ by definition. Consequently, if the change in the treatment rate from period 0 to 1 in the control group is not too large, $\left[B^{'}_{-}; B^{'}_{+}\right]$ will be narrow. Point identification can even be obtained if $\mathbb{P}(D_{t_1,g_c} = 1) = \mathbb{P}(D_{t_0,g_c} = 1)$.

3 Inference

The objective of this section is to build up confidence intervals (CI) for ATT_{t_1,g_t} based upon the identification results of section 2. I denote LB^{θ}_{x} and UB^{θ}_{x} the lower and upper bounds of the CI of a parameter θ with $x\%$ asymptotic coverage. A first candidate is $CI^1 = \left[LB^{\frac{DID}{DID^P}}_{(1-\alpha)}; UB^{\frac{DID}{DID^P}}_{(1-\alpha)}\right]$. In the no always takers special case, common trend is enough for CI^1 to be a consistent CI for ATT_{t_1,g_t}, since $ATT_{t_1,g_t} = \frac{DID}{DID^P}$. But when there are always takers, CI^1 is a CI for ATT_{t_1,g_t} (i.e. $ATT_{t_1,g_t} = \frac{DID}{DID^P}$) only under the very strong assumption that ATT do not vary across time × group cells. In such cases, partial identification results might allow deriving CI for ATT_{t_1,g_t} under weaker assumptions. This is the purpose of Proposition 3.

Proposition 3: CI for ATT_{t_1,g_t}based on partial identification results

i) Under A.1 and the supplementary assumption that $\exists (m, M) \in \mathbb{R}^2 / \mathbb{P}(m \leq Y(0) \leq M) = 1$, $CI^2 = \left[LB^{B-}_{(1-\alpha)}; UB^{B+}_{(1-\alpha)}\right]$ and $CI^3 = \left[LB^{B-}_{(1-2\alpha)}; UB^{B+}_{(1-2\alpha)}\right]$ are CI for ATT_{t_1,g_t} with asymptotic coverage of

$(1-\alpha)\%$.

ii) Under A.1 and the supplementary assumptions that $\exists(m, M) \in \mathbb{R}^2/\mathbb{P}(m \leq Y(0) \leq M) = 1$ *and that* $ATT_{t_1,g_c} = ATT_{t_0,g_c}$*, if either* $\mathbb{P}(D_{t_1,g_c} = 1) - \mathbb{P}(D_{t_0,g_c} = 1) \neq 0$ *or* $\mathbb{P}(D_{t_0,g_t} = 1) \neq 0$*, then* $CI^4 = \left[LB^{B'_-}_{(1-\alpha)}; UB^{B'_+}_{(1-\alpha)}\right]$ *and* $CI^5 = \left[LB^{B'_-}_{(1-2\alpha)}; UB^{B'_+}_{(1-2\alpha)}\right]$ *are CI for* ATT_{t_1,g_t} *with asymptotic coverage of* $(1-\alpha)\%$.

Based on the first partial identification result in Proposition 2, one can build a CI for ATT_{t_1,g_t} with $(1-\alpha)\%$ asymptotic coverage using the lower bound of the $(1-\alpha)\%$ CI of B_- and the upper bound of the $(1-\alpha)\%$ CI of B_+. This yields CI^2. As shown in Imbens and Manski [2004], using $(1-2\alpha)\%$ lower and upper bounds will also yield a CI for ATT_{t_1,g_t} with $(1-\alpha)\%$ asymptotic coverage. This is CI^3. However, it suffers from uniform convergence issues: when we get close to point identification ($\mathbb{P}_{AT} \to 0$), CI^3 will be narrower than CI^1 despite the fact that it is based on a partial identification result whereas CI^1 relies on point identification and stronger assumptions. To circumvent this issue, Imbens and Manski introduce a third CI lying in-between the $(1-\alpha)\%$ and the $(1-2\alpha)\%$ CI. It accounts for the fact that because the parameter is partially identified, the $(1-\alpha)\%$ CI is too conservative and also avoids the above mentioned uniform convergence issue. Stoye [2009] shows that this third CI relies on a superefficiency condition which is verified when by construction $\widehat{B_-} \leq \widehat{B_+}$ and when

$$\sqrt{n}\begin{pmatrix} \widehat{B_-} - B_- \\ \widehat{B_+} - B_+ \end{pmatrix} \xrightarrow{d} \mathcal{N}(0, \Sigma)$$

uniformly in $\mathcal{P}$. While the former is true here, the latter is not as shown in Proposition 4. Therefore, this third CI cannot be used here.

Finally, based on the second identification result in Proposition 2 which relies on stronger identifying assumptions, one can use B'_- and B'_+ instead of B_- and B_+ to build up CI for ATT_{t_1,g_t}. Using the lower bound of the $(1-\alpha)\%$ CI of B'_- and the upper bound of the $(1-\alpha)\%$ CI of B'_+ yields CI^4. Using the corresponding $(1-2\alpha)\%$ bounds yields CI^5.

Proposition 3 shows how to build up CI for ATT_{t_1,g_t} based upon CI for B_-, B_+, B'_- and B'_+. I show now how to construct such CI for B_- and B_+. Let $(Y_i,\, D_i,\, T_i,\, G_i)_{1\leq i\leq n}$ be an iid sample of size n drawn from the distribution of $(Y,\, D,\, T,\, G)$. I assume that $\mathbb{P}(T = i,\, G = j) > 0\ \forall(i,\, j) \in \{t_0; t_1\} \times \{g_c; g_t\}$ and that Y is bounded, meaning that $\exists(m,\, M) \in \mathbb{R}^2 / \mathbb{P}(m \leq Y(0) \leq M) = 1$, where m and M are known by the econometrician. Empirical counterparts are used to estimate B_- and B_+. I consider the asymptotic behavior of $\widehat{B_-}$ and $\widehat{B_+}$. On that purpose, I define a variance matrix $\Sigma = \begin{pmatrix} \sigma_1^2 & \rho \\ \rho & \sigma_2^2 \end{pmatrix}$ whose explicit expression is given in Appendix B and which can be consistently estimated by $\widehat{\Sigma}$.

Proposition 4: $\sqrt{n}$-consistency of $\widehat{B_-}$ and $\widehat{B_+}$.

If $B^0(M, m) > \mathbb{E}(Y_{t_1,g_t} | \, D = 1) - M$,

$$\sqrt{n}\left(\widehat{B_-} - B_-\right) \xrightarrow{d} \mathcal{N}(0, \sigma_1^2).$$

If $B^0(M, m) = \mathbb{E}(Y_{t_1,g_t} | \, D = 1) - M$,

$$\sqrt{n}\left(\widehat{B_-} - B_-\right) \xrightarrow{d} S^1$$

where $S^1 = max\left(N^1; N^2\right)$ *with* $\begin{pmatrix} N^1 & N^2 \end{pmatrix}' \sim \mathcal{N}(0, \Sigma)$.

If $B^0(M, m) < \mathbb{E}(Y_{t_1,g_t} | \, D = 1) - M$,

$$\sqrt{n}\left(\widehat{B_-} - B_-\right) \xrightarrow{d} \mathcal{N}(0, \sigma_2^2).$$

Similarly one can show that $\widehat{B_+}$ *is* $\sqrt{n}$*-consistent with three possible limiting distributions depending on the respective positions of* $B^0(m, M)$ *and* $\mathbb{E}(Y_{t_1,g_t} | \, D = 1) - m$.

B_- is not differentiable at $B^0(M, m) = \mathbb{E}(Y_{t_1,g_t} | \, D = 1) - M$ and B^+ is not differentiable at $B^0(m, M) = \mathbb{E}(Y_{t_1,g_t} | \, D = 1) - m$. Therefore, $\sqrt{n}\left(\widehat{B_-} - B_-\right)$ and $\sqrt{n}\left(\widehat{B_+} - B_+\right)$ do not converge to a normal distribution uniformly in $\mathcal{P}$. If $B^0(M, m) > \mathbb{E}(Y_{t_1,g_t} | \, D = 1) - M$, $\sqrt{n}\left(\widehat{B_-} - B_-\right)$ converges to a normal distribution. If $B^0(M, m) < \mathbb{E}(Y_{t_1,g_t} | \, D = 1) - M$, it converges to another normal distribution. If $B^0(M, m) = \mathbb{E}(Y_{t_1,g_t} | \, D = 1) - M$, its limiting distribution is non standard.

In all cases, it is possible to build CI for B_- and B_+. Let us consider B_- (the reasoning follows the same steps for B_+). If $B^0(M, m) > \mathbb{E}(Y_{t_1,g_t} | \, D = 1) - M$, a CI for B_- is $CI^A =$

$\left[\widehat{B^0(M,m)} - \frac{q_{1-\frac{\alpha}{2}}\times\widehat{\sigma_1^2}}{\sqrt{n}};\widehat{B^0(M,m)} + \frac{q_{1-\frac{\alpha}{2}}\times\widehat{\sigma_1^2}}{\sqrt{n}}\right]$, where $q_{1-\frac{\alpha}{2}}$ is the $1-\frac{\alpha}{2}^{th}$ quantile of a $\mathcal{N}(0,1)$ distribution. If $B^0(M,m) = \mathbb{E}(Y_{t_1,g_t}|\, D=1) - M$, a CI for B_- is $CI^B = \left[\widehat{B_-} + \frac{\widetilde{q}_{\frac{\alpha}{2}}}{\sqrt{n}};\widehat{B_-} + \frac{\widetilde{q}_{1-\frac{\alpha}{2}}}{\sqrt{n}}\right]$, where $\widetilde{q}_{\frac{\alpha}{2}}$ and $\widetilde{q}_{1-\frac{\alpha}{2}}$ are the $\frac{\alpha}{2}^{th}$ and $1-\frac{\alpha}{2}^{th}$ quantiles of $\widetilde{S^1} = max\left(N^1;N^2\right)$ with $\begin{pmatrix} N^1 & N^2 \end{pmatrix}^{'} \sim \mathcal{N}\left(0,\widehat{\Sigma}\right)$. Finally, if $B^0(M,m) < \mathbb{E}(Y_{t_1,g_t}|\, D=1) - M$, a CI for B_- is
$CI^C = \left[\widehat{\mathbb{E}}(Y_{t_1,g_t}|\, D=1) - M - \frac{q_{1-\frac{\alpha}{2}}\times\widehat{\sigma_2^2}}{\sqrt{n}};\widehat{\mathbb{E}}(Y_{t_1,g_t}|\, D=1) - M + \frac{q_{1-\frac{\alpha}{2}}\times\widehat{\sigma_2^2}}{\sqrt{n}}\right]$.

But $B^0(M,m)$ and $\mathbb{E}(Y_{t_1,g_t}|\, D=1) - M$ are unknown, hence the need to find CI with $(1-\alpha)\%$ asymptotic coverage irrespective of their respective position. This is achieved by choosing CI^A when $\widehat{B^0(M,m)}$ is more than $\frac{ln(n)}{\sqrt{n}}$ above $\widehat{\mathbb{E}}(Y_{t_1,g_t}|\, D=1) - M$, CI^B when $\widehat{B^0(M,m)}$ is less than $\frac{ln(n)}{\sqrt{n}}$ away from $\widehat{\mathbb{E}}(Y_{t_1,g_t}|\, D=1) - M$, and CI^C when $\widehat{B^0(M,m)}$ is more than $\frac{ln(n)}{\sqrt{n}}$ below $\widehat{\mathbb{E}}(Y_{t_1,g_t}|\, D=1) - M$.[4] The reason why this decision rule yields a CI with $(1-\alpha)\%$ asymptotic coverage uniformly in $B^0(M,m)$ and $\mathbb{E}(Y_{t_1,g_t}|\, D=1) - M$ is that since $\frac{1}{\sqrt{n}} = o\left(\frac{ln(n)}{\sqrt{n}}\right)$, the probability to pick the "wrong" CI converges to 0.

Proposition 5: CI for $\widehat{B_-}$ and $\widehat{B_+}$ with uniform asymptotic coverage

$$CI = CI^A \times 1_{\left\{\widehat{\mathbb{E}}(Y_{t_1,g_t}|\, D=1) - M + \frac{ln(n)}{\sqrt{n}} < \widehat{B^0(M,m)}\right\}}$$

$$+CI^B \times 1_{\left\{\widehat{\mathbb{E}}(Y_{t_1,g_t}|\, D=1) - M - \frac{ln(n)}{\sqrt{n}} \leq \widehat{B^0(M,m)} \leq \widehat{\mathbb{E}}(Y_{t_1,g_t}|\, D=1) - M + \frac{ln(n)}{\sqrt{n}}\right\}}$$

$$+CI^C \times 1_{\left\{\widehat{B^0(M,m)} + \frac{ln(n)}{\sqrt{n}} < \widehat{\mathbb{E}}(Y_{t_1,g_t}|\, D=1) - M\right\}}$$

is a CI for B_- with $(1-\alpha)\%$ asymptotic coverage uniformly in $B^0(M,m)$ and $\mathbb{E}(Y_{t_1,g_t}|\, D=1) - M$. A CI for B_+ with $(1-\alpha)\%$ asymptotic coverage uniformly in $B^0(m,M)$ and $\mathbb{E}(Y_{t_1,g_t}|\, D=1) - m$ can be constructed following the same steps.

Let us now consider $B_-^{'}$ and $B_+^{'}$. As in Proposition 4, one can show that whatever the value of $\mathbb{P}(D_{t_1,g_c}=1) - \mathbb{P}(D_{t_0,g_c}=1)$, $\mathbb{E}(Y_{t_1,g_c}|\, D=1) - \mathbb{E}(Y_{t_0,g_c}|\, D=1)$, $min(B^1\ ;\ B^2) - \mathbb{E}(Y_{t_1,g_t}|\, D=1) - M$ and $max(B^1\ ;\ B^2) - \mathbb{E}(Y_{t_1,g_t}|\, D=1) - m$, $\widehat{B_-^{'}}$ and $\widehat{B_+^{'}}$ are $\sqrt{n}$-consistent, with standard normal limiting distributions when those four quantities are different from 0, and with non standard limiting

[4] Instead of $ln(n)$, one can choose whatever sequence u_n such that $u_n \to +\infty$ and $\frac{u_n}{\sqrt{n}} \to 0$

distributions when one of them quantities is equal to 0. It is also possible to derive CI for B_- and B_+ with $(1-\alpha)\%$ asymptotic coverage irrespective of the value of those four unknown quantities. Because both B'_- and B'_+ are not differentiable at 3 points, careful analysis of their limiting distribution requires distinguishing 27 cases. Similarly, the construction of uniform CI for $B^{'}_-$ and $B^{'}_+$ involves 27 auxiliary CI. Due to a concern for brevity, the two corresponding propositions are not presented here.

4 Application to the impact of varenicline on smoking cessation.

4.1 Data and methods

I use the data base of French smoking cessation clinics participating in the "Consultation Dépendance Tabagique" program (hereafter referred to as CDT). This program started in 2001 and led to the progressive implementation of smoking cessation services nationwide. 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 first visit, treatments may be prescribed to patients (nicotine replacement therapies...). Follow-up visits are offered during which CO measures are usually made to validate tobacco abstinence.

Varenicline is a pharmacotherapy for smoking cessation support which was made available to these centers in February 2007. 59 services recorded at least one patient per year in 2006 and 2007 and followed at least 50% of their patients. The kernel density estimate of the rate of prescription of varenicline per center 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 services, less than 3% of all patients consulted have been prescribed varenicline during the year following its release. In 13 services, more than 20% of patients were prescribed varenicline. I exploit this to estimate the impact of varenicline on smoking cessation through a fuzzy DID identification strategy. The control group is made up of patients registered by the 15 "below 3% prescription rate" services, whereas the treatment group consists of patients recorded by "above 20% prescription rate" centers. Period 0 goes from February 2006 to January 2007, and period 1 from February 2007 to January 2008.

[Figure 1 inserted here]

8 581 patients consulted those 28 services 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, which I refer to as the included sample, I compute a point prevalence abstinence rate, that is to say the share of patients whose last follow-up CO determination was inferior or equal to 5 parts per million (ppm).

4.2 Results

In Table 1, I provide descriptive statistics on patients per group of centers and per period of time. Patients consulted in those cessation services 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]). 17% of them suffer from chronic obstructive pulmonary diseases (COPD) and more than 30% suffer from tobacco related diseases (lung cancer, COPD...). They have therefore been classified as "hardcore" addicts in the medical literature (Le Faou et al. [2005] and Le Faou et al. [2008]).

[Table 1 inserted here]

In period 0, the prescription rate of varenicline was equal to 0% in control centers and to 0.01% in treatment centers (varenicline was prescribed to 6 patients recorded in the last week of January 2007, that is to say right before the release of varenicline). In period 1, it was equal to 1.6% in control centers and to 38.2% in treatment centers. This sharp rise in varenicline prescription in treatment centers entailed a strong decrease in the prescription of other treatments such as nicotine patch. Finally, from period 0 to 1, the point prevalence abstinence rate increased (from 53.7% to 56.9%) in treatment centers, whereas it decreased (from 46.6% to 41.6%) in control centers. Among treatment patients prescribed varenicline in period 0, abstinence rate was equal to 50.0%. Among control patients prescribed varenicline in period 1, abstinence rate was equal to 58.3%. Applying the formulas of section 2, I compute that $\widehat{B_-} = 19.1\%$ (P-value = 0.008) and $\widehat{B_+} = 24.5\%$ (P-value = 0.001). Finally, $\widehat{\frac{DID}{DID^P}} = 22.7\%$ (P-value=0.003).

$\widehat{B^0(M,m)}$ is higher than $\widehat{\mathbb{E}}(Y_{t_1,g_t}|D=1)-1-\frac{ln(5299)}{\sqrt{5299}}$, and $\widehat{B^0(m,M)}+\frac{ln(5299)}{\sqrt{5299}}$ is lower than $\widehat{\mathbb{E}}(Y_{t_1,g_t}|D=1)$. Consequently, the CI to be used for B_- and B_+ are CI^A (see Proposition 5). Then, using Proposition 3, I construct 3 CI for ATT_{t_1,g_t}: $CI^1=\left[LB_{95}^{\widehat{\frac{DID}{\widehat{DID^P}}}};UB_{95}^{\widehat{\frac{DID}{\widehat{DID^P}}}}\right]=[7.8\%;37.5\%]$, $CI^2=\left[LB_{95}^{B_-};UB_{95}^{B_+}\right]=[5.0\%:38.6\%]$, $CI^3=\left[LB_{90}^{B_-};UB_{90}^{B_+}\right]=[7.3\%:36.3\%]$. The uniform convergence issue mentioned in Imbens and Manski [2004] shows up here since CI^3 is shorter than CI^1. But here even CI^2 is enough to infer the sign of ATT_{t_1,g_t}.

Point identification of ATT_{t_1,g_t} relies on a strong constant treatment effect assumption whereas identification of $[B_-;\,B_+]$ is obtained under much weaker assumptions. Moreover, inference on $\widehat{B_-}$ is sufficient to draw inference on the sign of ATT_{t_1,g_t}. Finally, even using CI^2, inference on $\widehat{B_-}$ and $\widehat{B_+}$ yields a 95% CI for ATT_{t_1,g_t} which is only slightly broader than the one obtained when drawing inference on $\widehat{\frac{DID}{\widehat{DID^P}}}$. Therefore, one might consider that here, the parameters which achieve the best trade-off between the accuracy of the information they deliver and the identifying assumptions on which they rely are $\widehat{B_-}$ and $\widehat{B_+}$ and not $\widehat{\frac{DID}{\widehat{DID^P}}}$.

4.3 Robustness checks

The only substantial assumption which is needed to identify $[B_-;\,B_+]$ is the common trend assumption. To "test" it, 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 the top panel of Table 2 along with their P-values. Only the 2006-2007 DID is significant, which gives some credit to the common trend assumption. I also compute 2006-2007 placebo DID on 9 patients' observable characteristics. They are also displayed in Table 2. This test is less conclusive since 2 DID out of 9 are significantly different from 0 at the 95% level. For instance, daily cigarettes smoked increased by 1.45 more among treatment centers than among control centers patients from 2006 to 2007. Similarly, the percentage of patients suffering from COPD increased by 4.4 percentage points more in treatment than in control services. This might cast some doubt on the validity of the common trend assumption. However, the P-value of the DID on the percentage of successful quits from 2006 to 2007 is still the lowest by far out of the 14 DID computed in Table 2. Moreover, high number of daily cigarettes smoked and COPD are predictors of unsuccessful quits. Since my fuzzy DID identification strategy does not correct for diverging trends on those variables, it might underestimate the true effect of varenicline.

Attrition seems orthogonal to the interaction of period 1 and treatment centers, since the DID computed on the percentage of patients included is low and insignificant (+2.2%, P-value = 0.30). Therefore, estimates do not seem contaminated by attrition bias. However, the delay between patients' first visit and the last CO measure available increased more in treatment than in control clinics. This is very likely to be because varenicline being a newly released drug with more severe secondary effects than nicotine patch, doctors put more effort in following their patients over a longer period of time to ensure they tolerate it well. Anyway, since smoking cessation is known to be a "duration" type of process, observing patients over a longer period of time in period 1 than in period 0 in treatment clinics can only bias downward my estimate.

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 services. I investigate the sensitivity of the results to the 3%-20% rule as a robustness check. I ran the same analysis with 9 different pairs of thresholds and always got $\widehat{B_-} \geq 0$ with 6 P-values lower than 0.05. The results of this last robustness check are displayed in the bottom panel of Table 2.

[Table 2 inserted here]

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

Finally, I investigate where the difference in varenicline prescription rates across services comes from. A first hypothesis is that this might be because services attend different type of patients with different needs in terms of drug prescription. Simple probit regressions of varenicline prescription on the 9 patients characteristics used above indicate that it is positively related to 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 services). Moreover, when the treatment clinics dummy is added to this probit regression model, the pseudo R-squared rises from 0.02 to 0.29, which indicates that differences in varenicline prescription across services cannot be explained only by the heterogeneity of patients attended.

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. Since no information on professionals working in smoking cessations clinics is available in the CDT data base, a survey was conducted to collect some information on them. Only 7 services (4 treatment and 3 control) out of the 28 included in the analysis answered to it. Still, the 4 treatment services recorded 1 612 patients over period 0 and 1, that is to say 64.5% of the "treatment" sample, and the 3 control services recorded 1 828 patients, that is to say 65.3% of the control sample.

Information on 29 professionals was collected, 20 working in the control services, 9 working in the treatment services. Since the number of patients recorded by each center varies a lot, and because inside a given clinic professionals do not dedicate the same amount of time to smoking cessation consultations, each of these 29 professionals is assigned a weight which proxies the percentage of patients in the sample consulted by her. Let C_i denote the number of patients recorded by clinic i, $D_{i,j}$ denote the number of days per week dedicated to smoking cessation consultations by professional j who works in clinic i. Let $C = \sum\limits_i C_i$ be the total number of patients in the sample and $D_i = \sum\limits_j D_{i,j}$ be the total number of days per week dedicated to smoking cessation consultations by professionals working in clinic i. Each professional is assigned the weight $w_{i,j} = \frac{C_i}{C} \times \frac{D_{i,j}}{D_i}$.

It appears that as per these weights, patients consulted in treatment clinics had a higher "probability" of being consulted by a doctor (76% against 47%, P-value = 0.12). On the contrary, they had a lower probability of being consulted by a psychologist (3% against 18%, P-value = 0.23) or by someone trained to behavioral and cognitive therapies (4% against 48%, P-value = 0.006). Finally, treatment services consulted 193 new patients per full time working professional in 2007, against only 75 in control services.

Contrarily to nicotine replacement therapies, varenicline must be prescribed by a doctor. This sharp difference in prescription rates across the two groups of centers 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 to be 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 services, lays more the emphasis on giving them intensive psychological support, hence the higher share of professionals trained to behavioral and cognitive therapies and the lower number of patients consulted per professional.

5 Summary and conclusions

This paper provides new identification results applying to fuzzy DID. Most of them hold under a common trend assumption on the outcome only, whereas the IV result commonly invoked in such settings holds under two common trends (on the outcome and on the treatment) and a monotonicity assumption. This single common trend assumption is sufficient to identify an ATT when there are no always takers, or at least its sign when there are "few" of them. When the shares of always takers are "large", supplementary assumptions must be taken. For instance, identification of an ATT can be obtained under the assumption that ATT do not vary across time and group. The milder assumption that the ATT in the control group did not change from period 0 to 1 substantially improves partial identification. This last result is of particular interest in applications considering the extension of a policy because in such situations it is likely to yield a narrow identification region.

I present an application in which the bounds I derive allow drawing inference on the sign of an ATT. This is because in this example, there are few always takers. Had there been more of them, the identification region would have been too large to infer the sign of the ATT. Consequently, in a fuzzy DID, common trend on Y is sufficient to obtain accurate information on the ATT when there are few always takers, even if there are many never takers. Conversely, in applications with many always takers, identification heavily relies on common trend on D and on monotonicity, or on alternative assumptions (constant ATT across time and group, constant ATT in the control group...).

Appendix A: multivariate treatment

Following Angrist and Imbens [1995], I assume that treatment is multivariate: $D \in \{0,1,...,K\}$. I define the corresponding $K+1$ potential outcomes of an individual: $Y(k),\ k \in \{0,1,...,K\}$. Only one outcome is observed, which is denoted

$$Y = Y(0) + (Y(1) - Y(0)) \times 1_{\{D=1\}} + ... + (Y(K) - Y(0)) \times 1_{\{D=K\}}$$

or alternatively

$$Y = Y(0) + (Y(1) - Y(0)) \times 1_{\{D \geq 1\}} + ... + (Y(K) - Y(K-1)) \times 1_{\{D \geq K\}}.$$

I denote

$$DID^P = \mathbb{P}(D_{t_1,g_t} > 0) - \mathbb{P}(D_{t_0,g_t} > 0) - [\mathbb{P}(D_{t_1,g_c} > 0) - \mathbb{P}(D_{t_0,g_c} > 0)]$$

and

$$DID^D = \mathbb{E}(D_{t_1,g_t}) - \mathbb{E}(D_{t_0,g_t}) - [\mathbb{E}(D_{t_1,g_c}) - \mathbb{E}(D_{t_0,g_c})].$$

I also define two types of parameters of interest:

$$ACE^1_{i,j} = \sum_{k=1}^{K} \mathbb{E}(Y_{i,j}(k) - Y_{i,j}(0) |\, D = k) \times \frac{\mathbb{P}(D_{i,j} = k)}{\mathbb{P}(D_{i,j} > 0)}$$

and

$$ACE^2_{i,j} = \sum_{k=1}^{K} \mathbb{E}(Y_{i,j}(k) - Y_{i,j}(k-1) |\, D \geq k) \times \frac{\mathbb{P}(D_{i,j} \geq k)}{\mathbb{E}(D_{i,j})}$$

which are weighted sums of ATT with weights summing up to 1.

Under those notations, it can be shown that

Lemma 1': Non identification.

Under A.1,

$$DID = \sum_{k=1}^{K} \mathbb{E}(Y_{t_1,g_t}(k) - Y_{t_1,g_t}(k-1) |\, D \geq k) \times \mathbb{P}(D_{t_1,g_t} \geq k) - \sum_{k=1}^{K} \mathbb{E}(Y_{t_0,g_t}(k) - Y_{t_0,g_t}(k-1) |\, D \geq k) \times \mathbb{P}(D_{t_0,g_t} \geq k)$$

$$-\sum_{k=1}^{K}\mathbb{E}(Y_{t_1,g_c}(k)-Y_{t_1,g_c}(k-1)|\, D\geq k)\times\mathbb{P}(D_{t_1,g_c}\geq k)+\sum_{k=1}^{K}\mathbb{E}(Y_{t_0,g_c}(k)-Y_{t_0,g_c}(k-1)|\, D\geq k)\times\mathbb{P}(D_{t_0,g_c}\geq k)$$

and

$$DID=\sum_{k=1}^{K}\mathbb{E}(Y_{t_1,g_t}(k)-Y_{t_1,g_t}(0)|\, D=k)\times\mathbb{P}(D_{t_1,g_t}=k)-\sum_{k=1}^{K}\mathbb{E}(Y_{t_0,g_t}(k)-Y_{t_0,g_t}(0)|\, D=k)\times\mathbb{P}(D_{t_0,g_t}=k)$$

$$-\sum_{k=1}^{K}\mathbb{E}(Y_{t_1,g_c}(k)-Y_{t_1,g_c}(0)|\, D=k)\times\mathbb{P}(D_{t_1,g_c}=k)+\sum_{k=1}^{K}\mathbb{E}(Y_{t_0,g_c}(k)-Y_{t_0,g_c}(0)|\, D=k)\times\mathbb{P}(D_{t_0,g_c}=k),$$

so that none of the parameters of interest is identified.

Proposition 1': Identification.

i) In the no always takers special case, A.1 is sufficient for $ACE^1_{t_1,g_t}$ and $ACE^2_{t_1,g_t}$ to be identified.

$$ACE^1_{t_1,g_t}=\frac{DID}{\mathbb{P}(D_{t_1,g_t}>0)}$$

and

$$ACE^2_{t_1,g_t}=\frac{DID}{\mathbb{E}(D_{t_1,g_t})}.$$

ii) Under A.1 and the supplementary assumption that

$\forall(i,j)\in\{t_0;t_1\}\times\{g_c;g_t\}\,,\forall k\in\{1,...,K\}\,,\mathbb{E}(Y_{i,j}(k)-Y_{i,j}(k-1)|D\geq k)=\mathbb{E}(Y_{t_1,g_t}(1)-Y_{t_1,g_t}(0)|D\geq 1)$

then

$$\forall(i,j)\in\{t_0;t_1\}\times\{g_c;g_t\}\,,\ ACE^2_{i,j}=\frac{DID}{DID^D}.$$

Before stating Proposition 2', I define the following function:

$$B^0(u,v)=\frac{DID+\big(\mathbb{E}(Y_{t_0,g_t}|\,D>0)-u\big)\times\mathbb{P}(D_{t_0,g_t}>0)+\big(\mathbb{E}(Y_{t_1,g_c}|\,D>0)-u\big)\times\mathbb{P}(D_{t_1,g_c}>0)-\big(\mathbb{E}(Y_{t_0,g_c}|\,D>0)-v\big)\times\mathbb{P}(D_{t_0,g_c}>0)}{\mathbb{P}(D_{t_1,g_t}=1)}.$$

Proposition 2': Partial identification.

i) Under A.1 and the supplementary assumption that $\exists(m,\,M)\in\mathbb{R}^2/\,\mathbb{P}(m\leq Y(0)\leq M)=1$,

$$B_-\leq ACE^1_{t_1,g_t}\leq B_+.$$

$B_-=max\left(B^0(M,m)\ ;\ \mathbb{E}(Y_{t_1,g_t}|\,D>0)-M\right)$ *and* $B_+=min\left(B^0(m,M)\ ;\ \mathbb{E}(Y_{t_1,g_t}|\,D>0)-m\right),$

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

$\mathbb{P}(D_{t_0,g_t} > 0) + \mathbb{P}(D_{t_1,g_c} > 0) + \mathbb{P}(D_{t_0,g_c} > 0) \leq \mathbb{P}(D_{t_1,g_t} > 0)$ *is a sufficient condition to have that at least one of the two bounds is informative.*

ii) Under A.1 and the supplementary assumptions that $\exists (m, M) \in \mathbb{R}^2 / \mathbb{P}(m \leq Y(0) \leq M) = 1$ *and that* $\forall k \in \{1, ..., K\}$, $\mathbb{E}(Y_{t_0,g_c}(k) - Y_{t_0,g_c}(0)|\, D = k) = \mathbb{E}(Y_{t_1,g_c}(k) - Y_{t_1,g_c}(0)|\, D = k)$,

$$B_-^{'} \leq ACE^1_{t_1,g_t} \leq B_+^{'}.$$

$B_-^{'} = max \left(\frac{DID + \left(\mathbb{E}(Y_{t_0,g_t}|\, D>0) - M\right) \times \mathbb{P}(D_{t_0,g_t}>0) + A^-}{\mathbb{P}(D_{t_1,g_t}>0)} \;;\; \mathbb{E}(Y_{t_1,g_t}|\, D > 0) - M \right)$,

with

$A^- = \sum_{k=1}^{K} \left((max\,(\mathbb{E}(Y_{t_1,g_c}|\, D = k); \mathbb{E}(Y_{t_0,g_c}|\, D = k)) - M) \times 1_{\left\{\mathbb{P}(D_{t_1,g_c}=k) - \mathbb{P}(D_{t_0,g_c}=k) > 0\right\}} \right) \times (\mathbb{P}(D_{t_1,g_c} = k) - \mathbb{P}(D_{t_0,g_c} = k))$
$+ \sum_{k=1}^{K} \left((min\,(\mathbb{E}(Y_{t_1,g_c}|\, D = k); \mathbb{E}(Y_{t_0,g_c}|\, D = k)) - m) \times 1_{\left\{\mathbb{P}(D_{t_1,g_c}=k) - \mathbb{P}(D_{t_0,g_c}=k) < 0\right\}} \right) \times (\mathbb{P}(D_{t_1,g_c} = k) - \mathbb{P}(D_{t_0,g_c} = k))$

and

$B_+^{'} = min \left(\frac{DID + \left(\mathbb{E}(Y_{t_0,g_t}|\, D>0) - m\right) \times \mathbb{P}(D_{t_0,g_t}>0) + A^+}{\mathbb{P}(D_{t_1,g_t}>0)} \;;\; \mathbb{E}(Y_{t_1,g_t}|\, D > 0) - m \right)$,

with

$A^+ = \sum_{k=1}^{K} \left((min\,(\mathbb{E}(Y_{t_1,g_c}|\, D = k); \mathbb{E}(Y_{t_0,g_c}|\, D = k)) - m) \times 1_{\left\{\mathbb{P}(D_{t_1,g_c}=k) - \mathbb{P}(D_{t_0,g_c}=k) > 0\right\}} \right) \times (\mathbb{P}(D_{t_1,g_c} = k) - \mathbb{P}(D_{t_0,g_c} = k))$
$+ \sum_{k=1}^{K} \left((max\,(\mathbb{E}(Y_{t_1,g_c}|\, D = k); \mathbb{E}(Y_{t_0,g_c}|\, D = k)) - M) \times 1_{\left\{\mathbb{P}(D_{t_1,g_c}=k) - \mathbb{P}(D_{t_0,g_c}=k) < 0\right\}} \right) \times (\mathbb{P}(D_{t_1,g_c} = k) - \mathbb{P}(D_{t_0,g_c} = k))$.

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

$\mathbb{P}(D_{t_0,g_t} > 0) + \sum_{k=1}^{K} |\mathbb{P}(D_{t_1,g_c} = k) - \mathbb{P}(D_{t_0,g_c} = k)| \leq \mathbb{P}(D_{t_1,g_t} > 0)$ *is a sufficient condition to have that at least one of the two bounds is informative.*

The only thing to be noted here is that the generalization of part ii) of Proposition 2 to multivariate treatment requires an assumption even stronger than in the case of binary treatment: not only should ATT be constant across time and group, but they should also be linear with respect to treatment doses $(\mathbb{E}(Y_{i,j}(k) - Y_{i,j}(k-1)|\, D \geq k) = \mathbb{E}(Y_{t_1,g_t}(1) - Y_{t_1,g_t}(0)|\, D \geq 1))$.

Appendix B: Explicit expression of Σ

Let

$X = \Big(\; YTG \quad Y(1-T)G \quad YT(1-G) \quad Y(1-T)(1-G) \quad YD(1-T)G \quad YDT(1-G) \quad YD(1-T)(1-G)$

$DTG \quad D(1-T)G \quad DT(1-G) \quad D(1-T)(1-G) \quad TG \quad (1-T)G \quad T(1-G) \quad (1-T)(1-G) \; \Big)'$

Let us denote $\theta = \mathbb{E}(X)$, $V = \mathbb{V}(X)$, $\widehat{\theta}$ the sample counterpart of θ and $\widehat{V}$ the sample counterpart of V.

Since Y is bounded, all the coordinates of X have a variance. Therefore, according to the central limit Theorem,

$\sqrt{n}(\widehat{\theta} - \theta) \xrightarrow{d} \mathcal{N}(0, V)$.

Let us denote

$$h(x) = \begin{pmatrix} \dfrac{\frac{x_1}{x_{12}} - \frac{x_2}{x_{13}} - \frac{x_3}{x_{14}} + \frac{x_4}{x_{15}} + \left(\frac{x_5}{x_9} - M\right) \times \frac{x_9}{x_{13}} + \left(\frac{x_6}{x_{10}} - M\right) \times \frac{x_{10}}{x_{14}} - \left(\frac{x_7}{x_{11}} - m\right) \times \frac{x_{11}}{x_{15}}}{\frac{x_8}{x_{12}}} \\ \frac{x_1}{x_{12}} - M \end{pmatrix},$$

which I define $\forall x = (x_1, x_2, x_3, x_4 x_5, x_6, x_7, x_8, x_9, x_{10}, x_{11}, x_{12}, x_{13}, x_{14}, x_{15}) \in \mathbb{R}^7 \times (\mathbb{R}^*)^8$.

$\theta \in \mathbb{R}^7 \times (\mathbb{R}^*)^8$ and h is continuously differentiable over $\mathbb{R}^7 \times (\mathbb{R}^*)^8$ with jacobian $H(x) \in M_{2,15}$.

I can therefore apply the delta method to state that:

$$\sqrt{n} \begin{pmatrix} \widehat{B^0(M, m)} - B^0(M, m) \\ \widehat{\mathbb{E}}(Y_{t_1,g_t} | \, D = 1) - M - (\mathbb{E}(Y_{t_1,g_t} | \, D = 1) - M) \end{pmatrix} \xrightarrow{d} \mathcal{N}(0; \Sigma)$$

where $\Sigma = H(\theta) V H(\theta)'$.

A consistent estimator of Σ is $\widehat{\Sigma} = H(\widehat{\theta}) \widehat{V} H(\widehat{\theta})'$.

Appendix C: proofs

Proof of Lemma 1:

$\forall (i, j) \in \{t_0; t_1\} \times \{g_c; g_t\}$, $Y_{i,j} = Y_{i,j}(1) \times D + Y_{i,j}(0) \times (1 - D) = (Y_{i,j}(1) - Y_{i,j}(0)) \times D + Y_{i,j}(0)$,

then,

$$DID = \mathbb{E}\left[(Y_{t_1,g_t}(1) - Y_{t_1,g_t}(0))D\right] - \mathbb{E}\left[(Y_{t_0,g_t}(1) - Y_{t_0,g_t}(0))D\right]$$

$$-\mathbb{E}\left[(Y_{t_1,g_c}(1)-Y_{t_1,g_c}(0))D\right]+\mathbb{E}\left[(Y_{t_0,g_c}(1)-Y_{t_0,g_c}(0))D\right]$$

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

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

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

$$-\left[\mathbb{E}(Y_{t_1,g_c}(1)-Y_{t_1,g_c}(0)|\,D=1)\times\mathbb{P}(D_{t_1,g_c}=1)-\mathbb{E}(Y_{t_0,g_c}(1)-Y_{t_0,g_c}(0)|\,D=1)\times\mathbb{P}(D_{t_0,g_c}=1)\right], \quad (2)$$

hence the result.

QED.

Proof of Proposition 1:

Proof of i)

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, (2) can be rewritten as

$$DID=\mathbb{E}(Y_{t_1,g_t}(1)-Y_{t_1,g_t}(0)|\,D=1)\times\mathbb{P}(D_{t_1,g_t}=1),$$

hence the result.

Proof of ii)

From (2),

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

If $\forall(i,j)\in\{t_0;t_1\}\times\{g_c;g_t\}$, $ATT_{i,j}=ATT$, then,

$$DID=ATT\times DID^P,$$

hence the result.

QED.

Proof of Proposition 2:

Proof of i)

Assume that $\exists (m, M) \in \mathbb{R}^2 / \mathbb{P}(m \leq Y(0) \leq M) = 1$. I denote

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

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

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

But for bounds to be sharp, the common trend assumption should hold, which implies:

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

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

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

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

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

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

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

Since $m \leq \mathbb{E}(Y_{t_1,g_t}(0)|\, D=1) \leq M$, this implies that we should have $A_2^- \leq A \leq A_2^+$, with

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

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

and

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

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

Consequently, we should have

$$max(A_1^-;A_2^-) \leq A \leq min(A_1^+;A_2^+). \tag{3}$$

Combining (2) and (3) and rearranging yields B_- and B_+, which are sharp by construction.

I show now that if none of the two bounds is informative then $\mathbb{P}_{AT} > \mathbb{P}(D_{t_1,g_t}=1)$. If B_- and B_+ are uninformative we have $B^0(M,m) < \mathbb{E}(Y_{t_1,g_t}|\, D=1)-M$ and $B^0(m,M) > \mathbb{E}(Y_{t_1,g_t}|\, D=1)-m$. Subtracting those two inequalities yields $\mathbb{P}_{AT} > \mathbb{P}(D_{t_1,g_t}=1)$. This implies that $\mathbb{P}_{AT} \leq \mathbb{P}(D_{t_1,g_t}=1)$ is a sufficient condition to have that at least one of the two bounds is informative.

To show that this condition is not sufficient to have that the two bounds are informative, it suffices to consider the following DGP. $M=1$, $m=0$, $\mathbb{P}(D_{t_1,g_t}=1)=1$, $\mathbb{P}(D_{t_0,g_t}=1)=\mathbb{P}(D_{t_1,g_c}=1)=0.1$, $\mathbb{P}(D_{t_0,g_c}=1)=0$, $\mathbb{E}(Y_{t_1,g_t}(1)|\, D=1)=\mathbb{E}(Y_{t_1,g_t}(0)|\, D=1)=1$, $\mathbb{E}(Y_{t_0,g_t}(0)|\, D=1)=\mathbb{E}(Y_{t_1,g_c}(0)|\, D=1)=0.5$, $\mathbb{E}(Y_{t_0,g_t}(0)|\, D=0)=\mathbb{E}(Y_{t_1,g_c}(0)|\, D=0)=1$, $\mathbb{E}(Y_{t_0,g_c}(0)|\, D=0)=0.9$. Those are all the quantities which are needed to compute B_- since the remaining expectations cancel out in the calculation. $\mathbb{P}_{AT}=0.2 \leq \mathbb{P}(D_{t_1,g_t}=1)=1$, the common trend assumption holds ($1\times1-0.5\times0.1-1\times0.9-0.5\times0.1-1\times0.9+0.9=0$), and B_- is not informative since it is equal to $\mathbb{E}(Y_{t_1,g_t}|\, D=1)-M$.

Proof of ii)

If $\exists (m,\, M) \in \mathbb{R}^2 / \, \mathbb{P}(m \leq Y(0) \leq M) = 1$,

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

and

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

If $ATT_{t_1,g_c} = ATT_{t_0,g_c} = ATT_{g_c}$, these two inequalities imply that

$$max\,(\mathbb{E}(Y_{t_1,g_c} |\, D = 1); \mathbb{E}(Y_{t_0,g_c} |\, D = 1)) - M \leq ATT_{g_c}$$

and

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

Moreover from (2) we get:

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

Therefore, combining this last equality with the two preceding inequalities yields B^1 and B^2 as lower or upper bounds to ATT_{t_1,g_t} depending on the sign of $\mathbb{P}(D_{t_1,g_c} = 1) - \mathbb{P}(D_{t_0,g_c} = 1)$. For some DGP, $min(B^1; B^2)$ might be smaller than $\mathbb{E}(Y_{t_1,g_t} |\, D = 1) - M$, which means that $min(B^1; B^2)$ is not a sharp lower bound, hence the need to set $B'_- = max\,\big(min(B^1; B^2); \mathbb{E}(Y_{t_1,g_t} |\, D = 1) - M\big)$ to ensure sharpness.

Finally, I show that $\mathbb{P}(D_{t_0,g_t} = 1) + |\mathbb{P}(D_{t_1,g_c} = 1) - \mathbb{P}(D_{t_0,g_c} = 1)| \leq \mathbb{P}(D_{t_1,g_t} = 1)$ is a sufficient condition to have that at least one of the two bounds is informative. Assume $\mathbb{P}(D_{t_1,g_c} = 1) - \mathbb{P}(D_{t_0,g_c} = 1) \geq 0$. None of the two bounds is informative if $B^1 < \mathbb{E}(Y_{t_1,g_t} |\, D = 1) - M$ and $B^2 > \mathbb{E}(Y_{t_1,g_t} |\, D = 1) - m$. Subtracting those two inequalities yields

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

$$+ (min\,(\mathbb{E}(Y_{t_1,g_c} |\, D = 1); \mathbb{E}(Y_{t_0,g_c} |\, D = 1)) - max\,(\mathbb{E}(Y_{t_1,g_c} |\, D = 1); \mathbb{E}(Y_{t_0,g_c} |\, D = 1))) \times (\mathbb{P}(D_{t_1,g_c} = 1) - \mathbb{P}(D_{t_0,g_c} = 1))$$

$$> (M - m) \times \mathbb{P}(D_{t_1,g_t} = 1) \tag{4}$$

Since (4) is a necessary condition to have that none of the bounds is informative, the converse inequality is sufficient to have that at least one of the two bounds is informative. But $\mathbb{P}(D_{t_0,g_t} = 1) + \mathbb{P}(D_{t_1,g_c} = 1) - \mathbb{P}(D_{t_0,g_c} = 1) \leq \mathbb{P}(D_{t_1,g_t} = 1)$ implies the converse inequality, hence the result. The proof is symmetric if $\mathbb{P}(D_{t_1,g_c} = 1) - \mathbb{P}(D_{t_0,g_c} = 1) < 0$.

QED.

Proof of Proposition 3:

Proof of i)

Under A.1 and the supplementary assumptions that $\exists(m,\, M) \in \mathbb{R}^2/\mathbb{P}(m \leq Y(0) \leq M) = 1$, $ATT_{t_1,g_t} \in [B_-; B_+]$ according to the first part of Proposition 2.

$$\underset{n\to+\infty}{lim} \mathbb{P}(ATT_{t_1,g_t} \geq LB^{B-}_{(1-\alpha)}) \geq \underset{n\to+\infty}{lim} \mathbb{P}(B_- \geq LB^{B-}_{(1-\alpha)}) = 1 - \frac{\alpha}{2}.$$

Similarly,

$$\underset{n\to+\infty}{lim} \mathbb{P}(ATT_{t_1,g_t} \leq UB^{B_+}_{(1-\alpha)}) \geq 1 - \frac{\alpha}{2}$$

which implies that

$$\underset{n\to+\infty}{lim} \mathbb{P}(LB^{B-}_{(1-\alpha)} \leq ATT_{t_1,g_t} \leq UB^{B_+}_{(1-\alpha)}) \geq 1 - \alpha.$$

Therefore, $CI^2 = \left[LB^{B-}_{(1-\alpha)}; UB^{B_+}_{(1-\alpha)}\right]$ is a CI for ATT_{t_1,g_t} with $(1-\alpha)\%$ asymptotic coverage.

Then, consider $\mathbb{P}(UB^{B-}_{(1-2\alpha)} \leq ATT_{t_1,g_t} \leq UB^{B_+}_{(1-2\alpha)})$.

If $ATT_{t_1,g_t} = B_-$,

$$\underset{n\to+\infty}{lim} \mathbb{P}(UB^{B-}_{(1-2\alpha)} \leq B_- \leq UB^{B_+}_{(1-2\alpha)}) = \underset{n\to+\infty}{lim} \mathbb{P}(UB^{B-}_{(1-2\alpha)} \leq B_-) - \underset{n\to+\infty}{lim} \mathbb{P}(B_- > UB^{B_+}_{(1-2\alpha)}) = 1 - \alpha$$

since the second term converges to 0.

If $ATT_{t_1,g_t} = B_+$, the same argument holds and $\underset{n\to+\infty}{lim} \mathbb{P}(UB^{B-}_{(1-2\alpha)} \leq ATT_{t_1,g_t} \leq UB^{B_+}_{(1-2\alpha)}) = 1 - \alpha$ as well.

If $B_- < ATT_{t_1,g_t} < B_+$,

$$\underset{n\to+\infty}{lim}\mathbb{P}(UB^{B_-}_{(1-2\alpha)} \leq ATT_{t_1,g_t} \leq UB^{B_+}_{(1-2\alpha)})$$

$$= \underset{n\to+\infty}{lim}\mathbb{P}(UB^{B_-}_{(1-2\alpha)} \leq ATT_{t_1,g_t}) - \underset{n\to+\infty}{lim}\mathbb{P}(ATT_{t_1,g_t} > UB^{B_+}_{(1-2\alpha)}) = 1.$$

Therefore, $\underset{n\to+\infty}{lim}\mathbb{P}(UB^{B_-}_{(1-2\alpha)} \leq ATT_{t_1,g_t} \leq UB^{B_+}_{(1-2\alpha)}) \geq 1-\alpha$ whatever the value of ATT_{t_1,g_t} so that $CI^3 = \left[LB^{B_-}_{(1-2\alpha)}; UB^{B_-}_{(1-2\alpha)}\right]$ is also a CI for ATT_{t_1,g_t} with $(1-\alpha)\%$ asymptotic coverage.

Proof of ii)

The proof follows the same steps as in i), once noted that under A.1 and the supplementary assumptions that $\exists(m,\, M) \in \mathbb{R}^2/\forall k \in \{0;1\}\ \mathbb{P}(m \leq Y(k) \leq M) = 1$ and that $ATT_{t_1,g_c} = ATT_{t_0,g_c}$, $ATT_{t_1,g_t} \in \left[B^{'}_-; B^{'}_+\right]$ as per the second part of Proposition 2.

QED.

Proof of Proposition 4:

By the delta method,

$$\sqrt{n}\left(\widehat{B^0(M,m)} - B^0(M,m)\right) \xrightarrow{d} \mathcal{N}(0;\sigma_1^2).$$

By the central limit theorem,

$$\sqrt{n}\left(\widehat{\mathbb{E}}(Y_{t_1,g_t}|\,D=1) - M - (\mathbb{E}(Y_{t_1,g_t}|\,D=1) - M)\right) \xrightarrow{d} \mathcal{N}(0;\sigma_2^2).$$

If $B^0(M,m) > \mathbb{E}(Y_{t_1,g_t}|\,D=1) - M$,

$$\sqrt{n}\left(\widehat{B_-} - B_-\right) = \sqrt{n}\left(max\left(\widehat{B^0(M,m)}; \widehat{\mathbb{E}}(Y_{t_1,g_t}|\,D=1) - M\right) - max\left(B^0(M,m); \mathbb{E}(Y_{t_1,g_t}|\,D=1) - M\right)\right)$$

$$= \sqrt{n}\left(\widehat{B^0(M,m)} - B^0(M,m)\right) + \sqrt{n}\left(max\left(\widehat{B^0(M,m)}; \widehat{\mathbb{E}}(Y_{t_1,g_t}|\,D=1) - M\right) - \widehat{B^0(M,m)}\right).$$

The second term is $o_p(1)$ because $max\left(\widehat{B^0(M,m)}; \widehat{\mathbb{E}}(Y_{t_1,g_t}|\,D=1) - M\right) = \widehat{B^0(M,m)}$ with probability approaching 1. This implies the result.

If $B^0(M,m) < \mathbb{E}(Y_{t_1,g_t}|\,D=1) - M$, the proof is symmetric.

If $B^0(M,m) = \mathbb{E}(Y_{t_1,g_t}|\, D=1) - M$,

$$\sqrt{N}\left(\widehat{B_-} - B_-\right) = max\left(\sqrt{N}\left(\widehat{B^0(M,m)} - B^0(M,m)\right); \sqrt{N}\left(\widehat{\mathbb{E}}(Y_{t_1,g_t}|\, D=1) - M - (\mathbb{E}(Y_{t_1,g_t}|\, D=1) - M)\right)\right).$$

Due to the continuous mapping Theorem,

$$max\left(\sqrt{N}\left(\widehat{B^0(M,m)} - B^0(M,m)\right); \sqrt{N}\left(\widehat{\mathbb{E}}(Y_{t_1,g_t}|\, D=1) - M - (\mathbb{E}(Y_{t_1,g_t}|\, D=1) - M)\right)\right) \hookrightarrow S^1 = \left(max\left(N^1; N^2\right)\right)$$

where $\begin{pmatrix} N_1 & N_2 \end{pmatrix}' \sim \mathcal{N}(0, \Sigma)$.

QED.

Proof of Proposition 5:

$$\mathbb{P}(B_- \in CI) = \mathbb{P}\left(B_- \in CI^A \mid \widehat{\mathbb{E}}(Y_{t_1,g_t}|\, D=1) - M + \frac{ln(n)}{\sqrt{n}} < \widehat{B^0(M,m)}\right) \times \mathbb{P}\left(\widehat{\mathbb{E}}(Y_{t_1,g_t}|\, D=1) - M + \frac{ln(n)}{\sqrt{n}} < \widehat{B^0(M,m)}\right)$$

$$+\mathbb{P}\left(B_- \in CI^B \mid \widehat{\mathbb{E}}(Y_{t_1,g_t}|\, D=1) - M - \frac{ln(n)}{\sqrt{n}} \leq \widehat{B^0(M,m)} \leq \widehat{\mathbb{E}}(Y_{t_1,g_t}|\, D=1) - M + \frac{ln(n)}{\sqrt{n}}\right)$$

$$\times \mathbb{P}\left(\widehat{\mathbb{E}}(Y_{t_1,g_t}|\, D=1) - M - \frac{ln(n)}{\sqrt{n}} \leq \widehat{B^0(M,m)} \leq \widehat{\mathbb{E}}(Y_{t_1,g_t}|\, D=1) - M + \frac{ln(n)}{\sqrt{n}}\right)$$

$$+\mathbb{P}\left(B_- \in CI^C \mid \widehat{B^0(M,m)} + \frac{ln(n)}{\sqrt{n}} < \widehat{\mathbb{E}}(Y_{t_1,g_t}|\, D=1) - M\right) \times \mathbb{P}\left(\widehat{B^0(M,m)} + \frac{ln(n)}{\sqrt{n}} < \widehat{\mathbb{E}}(Y_{t_1,g_t}|\, D=1) - M\right).$$

If $B^0(M,m) > \mathbb{E}(Y_{t_1,g_t}|\, D=1) - M$,

$$\mathbb{P}\left(\widehat{\mathbb{E}}(Y_{t_1,g_t}|\, D=1) - M + \frac{ln(n)}{\sqrt{n}} < \widehat{B^0(M,m)}\right)$$

$$= \mathbb{P}\left(\sqrt{n}\left(\left[\widehat{B^0(M,m)} - B^0(M,m)\right] - \left[\widehat{\mathbb{E}}(Y_{t_1,g_t}|D=1) - \mathbb{E}(Y_{t_1,g_t}|D=1)\right]\right) > ln(n) - \left(B^0(M,m) - (\mathbb{E}(Y_{t_1,g_t}|D=1) - M)\right)\sqrt{n}\right)$$

I denote V_n this sequence.

$$\underset{n\to+\infty}{lim} ln(n) - \left(B^0(M,m) - (\mathbb{E}(Y_{t_1,g_t}|\, D=1) - M)\right)\sqrt{n} = -\infty.$$

Consequently, $\forall x \in \mathbb{R}$, $\exists n_0 \in \mathbb{N}/n \geq n_0 \Rightarrow$

$$\mathbb{P}\left(\sqrt{n}\left(\widehat{B^0(M,m)} - B^0(M,m)\right) - \sqrt{n}\left(\widehat{\mathbb{E}}(Y_{t_1,g_t}|\, D=1) - M - (\mathbb{E}(Y_{t_1,g_t}|\, D=1) - M)\right) > x\right) \leq V_n$$

Therefore,

$$\underset{n\to+\infty}{lim}\mathbb{P}\left(\sqrt{n}\left(\widehat{B^0(M,m)}-B^0(M,m)\right)-\sqrt{n}\left(\widehat{\mathbb{E}}(Y_{t_1,g_t}|\,D=1)-M-(\mathbb{E}(Y_{t_1,g_t}|\,D=1)-M)\right)>x\right)\leq\underset{n\to+\infty}{lim}V_n$$

A delta method and the central limit theorem imply that

$$\underset{n\to+\infty}{lim}\mathbb{P}\left(\sqrt{n}\left(\widehat{B^0(M,m)}-B^0(M,m)\right)-\sqrt{n}\left(\widehat{\mathbb{E}}(Y_{t_1,g_t}|\,D=1)-M-(\mathbb{E}(Y_{t_1,g_t}|\,D=1)-M)\right)>x\right)=1-F(x),$$

where $F(.)$ is the cdf of a random variable with a normal distribution.

Since this holds $\forall x\in\mathbb{R}$, we can let x go to $-\infty$ which yields $1\leq\underset{n\to+\infty}{lim}V_n$. Therefore,

$$\underset{n\to+\infty}{lim}\mathbb{P}\left(\widehat{\mathbb{E}}(Y_{t_1,g_t}|\,D=1)-M+\frac{ln(n)}{\sqrt{n}}<\widehat{B^0(M,m)}\right)=1,$$

which implies that

$$\underset{n\to+\infty}{lim}\mathbb{P}\left(B_-\in CI^B\mid\widehat{\mathbb{E}}(Y_{t_1,g_t}|\,D=1)-M-\frac{ln(n)}{\sqrt{n}}\leq\widehat{B^0(M,m)}\leq\widehat{\mathbb{E}}(Y_{t_1,g_t}|\,D=1)-M+\frac{ln(n)}{\sqrt{n}}\right)=0$$

and

$$\underset{n\to+\infty}{lim}\mathbb{P}\left(\widehat{B^0(M,m)}+\frac{ln(n)}{\sqrt{n}}<\widehat{\mathbb{E}}(Y_{t_1,g_t}|\,D=1)-M\right)=0.$$

Consequently,

$$\underset{n\to+\infty}{lim}\mathbb{P}(B_-\in CI)=\underset{n\to+\infty}{lim}\mathbb{P}\left(B^0(M,m)\in CI^A\right)=1-\alpha.$$

If $B^0(M,m)=\mathbb{E}(Y_{t_1,g_t}|\,D=1)-M$ or $B^0(M,m)<\mathbb{E}(Y_{t_1,g_t}|\,D=1)-M$, the same type of reasoning yields $\underset{n\to+\infty}{lim}\mathbb{P}(B_-\in CI)=1-\alpha$ which completes the proof.

QED.

References

[1] Abadie A. 2005. Semiparametric Difference-in-Differences Estimators. The Review of Economic Studies. 72(1), 1-19.

[2] Angrist J, Imbens G. 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.

[3] Angrist J, Imbens G, Rubin D. 1996. Identification of Causal Effects Using Instrumental Variables. Journal of the American Statistical Association. 91(434), 444- 455.

[4] Ashenfelter O, Card D. 1985. Using the Longitudinal Structure of Earnings to Estimate the Effect of Training Programs. The Review of Economics and Statistics, 67(4) 648-60.

[5] Bach L. 2010. Are Small-and-medium-sized Firms Really Credit Constrained ? Evidence from a French Targeted Credit Program. Working paper.

[6] 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.

[7] Beck F, Guilbert P, Gautier A. 2007. Baromètre Santé [French Health Barometer], Saint-Denis, INPES.

[8] Bertrand M, Duflo E, Mullainathan S. 2004. How Much Should We Trust Differences-in-Differences Estimates? The Quarterly Journal of Economics. 119(1), 249-275.

[9] Bleakley H, Chin A. 2004. Language Skills and Earnings: Evidence from Childhood Immigrants. The Review of Economics and Statistics. 86(2), 481-496.

[10] de Chaisemartin C. 2010. Instrumented difference in differences. Working paper.

[11] Duflo E. 2001. Schooling and Labor Market Consequences of School Construction in Indonesia: Evidence from an Unusual Policy Experiment. The American Economic Review. 91(4), 795-813.

[12] Evans W, Ringel J. 1999. Can higher cigarette taxes improve birth outcomes? Journal of Public Economics. 72 (1) 135–154.

[13] Field E. 2005. Property Rights and Investment in Urban Slums. Journal of the European Economic Association. 3(2-3), 279-290.

[14] Horowitz J, and Manski C. 1995. Identification and Robustness with Contaminated and Corrupted Data. Econometrica. 63(2), 281–302.

[15] Imbens G, Angrist J. 1994. Identification and Estimation of Local Average Treatment Effects. Econometrica. 62(2), 467-475.

[16] Imbens G, Manski C. 2004. Confidence intervals for partially identified parameters. Econometrica. 72(6), 1845–1857.

[17] Lee D. 2009. Training, Wages, and Sample Selection: Estimating Sharp Bounds on Treatment Effects. The Review of Economic Studies. 76(3), 1071–1102.

[18] Le Faou AL, Scemama O, Ruelland A, Menard J. 2005. [Characteristics of smokers seeking smoking cessation services: the CDT program]. Revue des Maladies Respiratoires. 22, 739-50.

[19] Le Faou AL, Baha M, Rodon N, Lagrue G, Menard J. 2009. Trends in the profile of smokers registered in a national database from 2001 to 2006: changes in smoking habits. Public Health. 123, 6-11.

[20] Manski C. 1990. Nonparametric Bounds on Treatment Effects. The American Economic Review. 80(2), 319-323. Papers and Proceedings of the Hundred and Second Annual Meeting of the American Economic Association.

[21] Small D, Tan Z. 2007. A Stochastic Monotonicity Assumption for the Instrumental Variables Method. Working Paper, Department of Statistics, Wharton School, University of Pennsylvania.

[22] Stoye J. 2009. More on confidence intervals for partially identified parameters. Econometrica. Econometrica. 77(4), 1299–131

Tables

Table 1: Descriptive Statistics

	Whole sample	**Test Centers**			**Control Centers**		
		2006	2007	P-value	2006	2007	P-value
Patients' characteristics							
% males	48.8%	47.9%	47.9%	0.98	48.5%	50.4%	0.30
Age	44.1	44.6	43.7	0.08	44.0	44.3	0.52
% employed	67.3%	65.3%	68.3%	0.11	65.3%	69.8%	0.01
% with no degree	17.0%	19.2%	21.0%	0.25	14.2%	14.1%	0.98
Daily cigarettes smoked	21.6	21.7	21.93	0.60	22.1	20.9	<0.01
FTND	5.9	5.8	5.8	0.29	6.0	5.9	0.11
% with AHAD>=11	39.8%	40.3%	39.1%	0.54	42.2%	37.7%	0.01
% with DHAD>=11	11.9%	13.1%	11.7%	0.29	11.6%	11.2%	0.72
% with chronic obstructive pulmonary diseases	16.7%	16.2%	18.1%	0.19	17.5%	15.1%	0.09
Treatment prescribed							
% prescribed nicotine patch	53.4%	75.0%	45.5%	<0.001	45.9%	49.7%	0.05
% prescribed varenicline	10.0%	0.01%	38.2%	<0.001	0%	1.6%	<0.001
Cessation Outcome							
Number of days between the first visit and the last CO measure	86.7	89.3	96.7	0.05	84.8	77.6	0.03
% of successful quits	49.3%	53.7%	56.9%	0.11	46.6%	41.6%	<0.01
N	5 299	1 195	1 303		1 300	1 501	

[1] FTND stands for Fagerström Test for Nicotine Dependence and is a measure of patients' degree of addiction.

[2] AHAD is the anxiety scale in the Hospital Anxiety Depression (HAD) scale, scored from 0 to 21, which is used to identify individuals with anxio-depressive disorders, with a threshold score of 11 (see Zigmond et al. [1983]).

[3] DHAD is the depression scale in the Hospital Anxiety Depression (HAD) scale, scored from 0 to 21, which is used to identify individuals with anxio-depressive disorders, with a threshold score of 11 (see Zigmond et al. [1983]).

[4] CO stands for carbon monoxide which is a biomarker for tobacco use.

Table 2: Robustness Checks

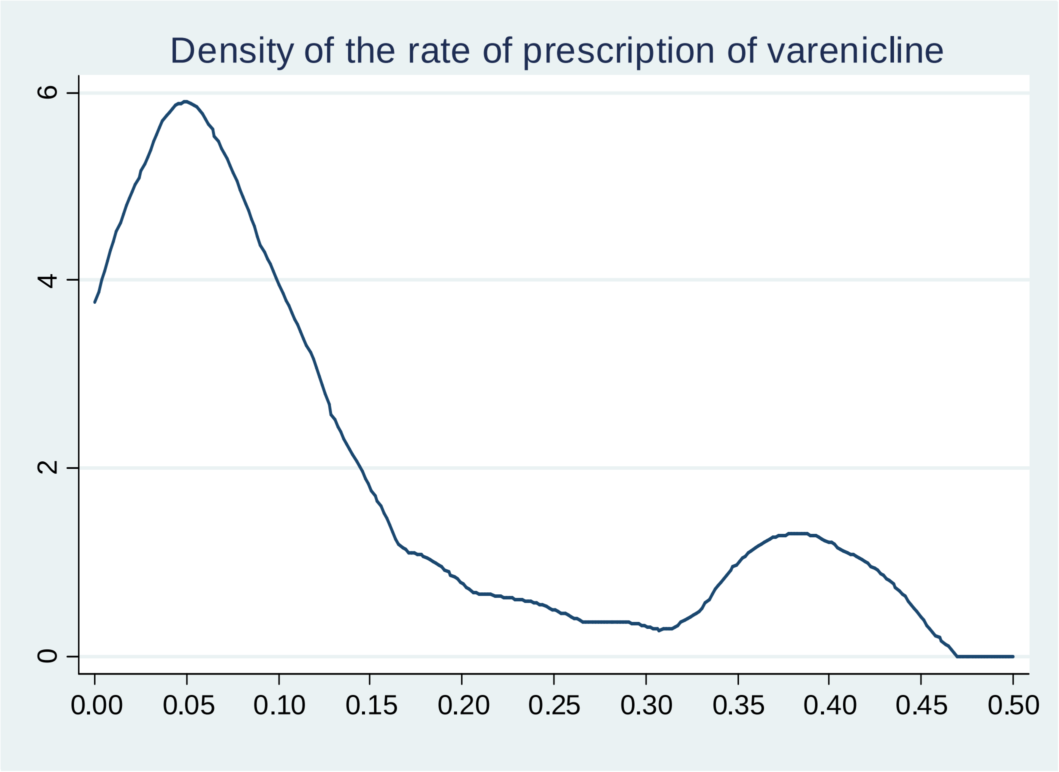
	Common Trend		
	Diff in diff	P-value	N
2003-2004	0.045	0.36	1 580
2004-2005	0.032	0.46	2 499
2005-2006	0.042	0.19	4 136
2006-2007	0.082	0.003	5 299
2007-2008	-0.043	0.17	4 400

	Placebo DID		
	Diff in diff	P-value	N
Patients' observable characteristics			
% Males	-0.020	0.46	5 299
Age	-1.153	0.08	5 298
% employed	-0.015	0.57	5 299
% with no degree	0.019	0.36	5 299
Daily cigarettes smoked	1.454	0.02	5 299
FTND	0.237	0.06	5 299
% with AHAD>=11	0.033	0.22	5 299
% with DHAD>=11	-0.010	0.59	5 299
% with chronic obstructive pulmonary diseases	0.043	0.04	5 299
Measurement of smoking status			
Number of days between the first visit and the last CO measure	14.653	0.004	5 299
% included	0.022	0.30	8 581

		P-value of B_ according to inclusion threshold		
		Test centers thresholds		
		Threshold 1: 15%	Threshold 2: 20%	Threshold 3: 25%
Control centers thresholds	Threshold 1: 2%	0.04	0.03	0.04
	Threshold 2: 3%	0.01	0.01	0.02
	Threshold 3: 4%	0.11	0.08	0.14

Figures

Figure 1: Density of the prescription rate of varenicline


Density of the rate of prescription of varenicline