

NONCOMPLIANCE BIAS CORRECTION BASED ON COVARIATES IN RANDOMIZED EXPERIMENTS

YVES ATCHADE *AND LEONARD WANTCHEKON†

November 1, 2005

Abstract

We propose some practical solutions for causal effects estimation when compliance to assignments is only partial and some of the standard assumptions do not hold. We follow the potential outcome approach but in contrast to Imbens and Rubin (1997), we require no prior classification of the compliance behaviour. When noncompliance is not ignorable, it is known that adjusting for arbitrary covariates can actually increase the estimation bias. We propose an approach where a covariate is adjusted for only when the estimate of the selection bias of the experiment as provided by that covariate is consistent with the data and prior information on the study. Next, we investigate cases when the overlap assumption does not hold and, on the basis of their covariates, some units are excluded from the experiment or equivalently, never comply with their assignments. In that context, we show that a consistent estimation of the causal effect of the treatment is possible based on a regression model estimation of the conditional expectation of the outcome given the covariates. We illustrate the methodology with several examples such as the access to influenza vaccine experiment (McDonald et al (1992) and the PROGRESA experiment (Shultz (2004)).

Key words: *Causal inference, Intent-to-treat, Noncompliance, Randomized experiments*

JEL Classification: *C14, C21, C52*

*Department of Mathematics and Statistics, University of Ottawa.

†Departments of Politics and Economics, New York University.

‡The authors are grateful to Dr McDonald for sharing with us the data from their study on the effect of influenza vaccination on morbidity. This work is supported in part by the Natural Sciences and Engineering Research Council of Canada.

1 Introduction

Randomized experiments are a widely accepted approach to infer causal relations in statistics and social sciences. The idea dates back at least to Neyman (1923) and Fisher (1935) and has been extended by D. Rubin and co-authors (Rubin (1974), Rubin (1978), Rosebaum and Rubin (1983)) to observational studies and other more general experimental designs. In this approach, causality is defined in terms of potential outcomes. The causal effect of a treatment, say Treatment 1 (compared to another treatment, Treatment 0) on the variable Y and on the statistical unit i is defined as $Y_i(1) - Y_i(0)$ where $Y_i(j)$ is the value we would observe on unit i if it receives Treatment j . The estimation of this effect is problematic because unit i cannot be given both treatments. This difficulty is circumvented by randomizing the receipt of the treatments. To estimate the causal effect of a treatment, two random samples of units are selected, the first group is assigned to Treatment 0 and the second group to Treatment 1. The difference in the sample means of Y (or some statistic of interest) over the two groups is used as an estimate of the causal effect of the treatment. The main idea is that randomization eliminates (at least in theory) any systematic difference between the two samples. We refer the reader to Holland (1986) for an interesting discussion of causality in statistics.

But in practice, it happens often that some units do not comply with their assignments. On a large scale, noncompliance limits the internal validity of the experiment and may introduce severe bias in the estimation of the causal effect of the experiment. A trivial example of the effect of noncompliance is one in which a harmful drug is administered and units assigned to the drug who recover after the experiment are merely those who did not comply with their assignment. In general, in presence of noncompliance, the causal effect of the treatment is more difficult to define (and less useful) as it depends on the treatment assigned to.

One widely used approach to deal with noncompliance is Intention-to-Treat (ITT) analysis (see e.g. Lee et al (1991)). ITT proposes precisely to ignore the noncompliance problem and to use the effect of the assignment as a proxy for the treatment effect. In ITT, one sees the noncompliance behaviour as part of the response to the treatment. Therefore, an ITT estimate can be interpreted as the “use-effectiveness” of the treatment. It is particularly useful when the noncompliance behaviour is believed to be stable over time. However, the ITT estimates can be strikingly different from the real treatment effect particularly when noncompliance is severe and/or changes over time. This could be a problem if the experimenter is mainly interested in the real

effect of the treatment.

An alternative to the Intent-to-Treat estimate is the Local Average Treatment Effect (LATE) proposed by Angrist et al (1996) (see also Imbens and Rubin (1997)). The LATE estimate is the ITT estimate divided by the proportion of units that comply with their assignment. Assuming four patterns of compliance behavior (compliers, never-takers, always-takers and defiers), these authors have shown (under some additional assumptions) that the LATE estimate can be given a causal interpretation as the causal effect of the treatment over the sub-population of compliers. The LATE estimate is interesting because compliers are precisely the units for which speaking of the causal effect of the treatment makes sense. But the approach is of a limited applicability since this sub-population of compliers is (generally) not known and cannot be defined precisely. Moreover, assuming such a rigid classification of compliance behavior may limit the external validity of the method.

In this paper, we take the approach that noncompliance is a random behavior driven by covariates and that there is no rigid classification of the compliance behavior. In order to define precisely the causal effect of the treatment we make the so-called exclusion restriction assumption (Angrist et al (2002)), assuming that $Y(t, d)$ and $Y(t', d)$ have the same distribution for any $t, t', d \in \{0, 1\}$. This amounts to assuming that the assignment has no direct effect on the outcome but through the received treatment. In this framework, we ask two questions. Is it possible to consistently estimate the causal effect of the treatment if:

1. the noncompliance behavior is not ignorable given the observed covariate?
2. some sub-population (called in the sequel “taboo population”) can never be enticed to take the treatments?

Regarding question 1 we show that when noncompliance is not ignorable given the available covariates, controlling for arbitrary covariates can actually worsen the estimation bias. We then propose an approach where prior information on the study is combined with the data to determine which covariates one should adjust for. The idea is based on the fact that adjusting for a covariate corrects only for the selection bias of the experiment as seen through that covariate (that is, after conditioning on that covariate). Therefore adjusting for a covariate can increase the bias if the covariate has a wrong picture of the selection bias of the experiment. When appropriate information is available, it is possible to find how well a covariate will estimate the selection bias of the experiment and whether that covariate should be adjusted for or not.

Question 2 is an example where the so-called overlapping assumption breaks down. It is quite frequent in medical studies where units are screened based on their covariates before partici-

pating in the experiments and yet the experimenter would like to estimate the causal effect of the treatment over the whole population (See also Wantchekon (2003) for a randomized experiment in political science where this problem arises¹). In this paper, we formalize the selection of the units in this type of experiments as a sequence of negative binomial trials and we show that if a (global) representation of the conditional expectation of the outcome given the covariates exists then it can be consistently estimated from the data and used to estimate the causal treatment over the whole population (including the taboo sub-population). By estimating the potential treatment effect on the untreatable units, that is, units that cannot be assigned neither to Treatment 1 nor to Treatment 0, our method provides a general and formal test of external validity of randomized experiments. This could turn out to be a very useful exercise, especially in medical trials where the experimenter is interested estimating the effect of a drug on a sub-population that cannot participate in the experiment (presumably because of potential side effects of the drug).

The rest of the paper is organized as follows. In Section 2, we develop the basic framework for covariate-based bias correction. This classical material is presented here for completeness. Its presentation also allows us to introduce the estimators that we use next. In Section 3.1, we present our new method for bias reduction in the presence of unobserved covariates and in Section 3.2, we show how to consistently estimate the causal treatment effect in presence of a taboo sub-population. We present several examples to illustrate our methods in Section 4. We then present a re-analysis of an experiment to assess the effect of an influenza vaccine conducted by McDonald et al (1992) (see also Hirano et al (2000)).

2 Noncompliance bias correction with observed covariates

In this section, we present the basic framework of the potential outcome model for causal inference in randomized experiments with noncompliance. We include these well known results here to make it easier for the reader to follow our discussion in Section 3. The basic assumptions are detailed in Section 2.1. Some consistent estimators of the causal effect are presented in Section 2.2.

¹Wantchekon (2003) presents results from an experiment in which *non-competitive* electoral districts were randomly assigned to “purified” national public goods and redistributive platforms by candidates competing in the 2001 presidential elections in Benin. We may want to estimate the potential causal effect of the treatment if *competitive* and ethnically diverse districts have participated in the experiment.

2.1 The basic framework

A random sample of n units from a reference population is taken and each unit is assigned to Treatment 1 or Treatment 0. Let (T_1, \dots, T_n) be the random vector of assignments; $T_i = j$ if unit i is assigned to Treatment j , $j = 0, 1$ and $i = 1, \dots, n$. Define the random variable $C_i = 1$ if unit i complies with its assignment and 0 if not. Finally define D_i to be the treatment actually received by unit i . $D_i = 1$ if unit i actually receives Treatment 1 and $D_i = 0$ if unit i actually receives Treatment 0. If $T_i = t$ and $D_i = d$, we observe the outcome $Y_i(t, d)$. We assume that $Y_i(t, d)$ is a random variable with finite expectation. Let X_i be a vector of covariates (pre-treatment variables) observed on unit i before its assignment to a treatment. We assume that $(Y_i(0, 0), Y_i(0, 1), Y_i(1, 0), Y_i(1, 1), X_i)_{1 \leq i \leq n}$ are independent and identically distributed (i.i.d.) and when there is no ambiguity, we shall omit the subscript i when referring to an arbitrary element of the sample. Following Angrist et al (1996) we assume that:

Assumption **(A1)**:

$$\Pr(Y_i(0, d) \in \cdot | X) = \Pr(Y_i(1, d) \in \cdot | X), \text{ a.s. for } d = 0, 1. \quad (1)$$

The idea behind (A1) is that the assignment variable T can affect the outcome variable Y only through the received treatment variable D so that the conditional distribution of $Y(0, d)$ and $Y(1, d)$ given X are the same for $d = 0, 1$. We find this assumption quite realistic in most applications. But it can be questionable in situations where assignment to a given treatment can significantly change the behavior of the subject and therefore have a direct impact on the outcome.

For $d = 0, 1$, let $\mu_d(x) := \mathbb{E}(Y(0, d)|X = x) = \mathbb{E}(Y(1, d)|X = x)$ and $\mu_d := \mathbb{E}(Y(0, d)) = \mathbb{E}(Y(1, d))$ where the second equalities follow from (A1). We are interested in the causal effect of Treatment 1 versus Treatment 0 at covariate level $X = x$ defined as

$$\tau(x) = \mu_1(x) - \mu_0(x), \quad (2)$$

and most importantly, the causal effect of the treatment over the population defined as:

$$\tau_p = \mu_1 - \mu_0. \quad (3)$$

The well-known problem in estimating $\tau(x)$ and τ_p is that all the outcome variables $Y_i(t, d)$ cannot be observed in the same time. Throughout the paper, we assume that the assignment is ignorable given the covariates; we assume that:

Assumption **(A2)**:

$$(Y_i(0, 0), Y_i(0, 1), Y_i(1, 0), Y_i(1, 1)) \perp T_i | X_i, \quad 1 \leq i \leq n, \quad (4)$$

where for random objects X, Y the notation $X \perp Y$ means that X and Y are independent and for random variables X, Y, Z , we write $X \perp Y|Z$ to say that X and Y are conditionally independent given Z . If we have $Cov(X, Y|Z) := E(XY|Z) - E(X|Z)E(Y|Z) = 0$ we shall say that X and Y are uncorrelated given Z . Similarly, if $Cov(X, Y) := E(XY) - E(X)E(Y) = 0$, we say that X and Y are uncorrelated.

Remark 1 1. *Assumption (A2) is not restrictive in randomized experiments. Often the assignments are actually randomized independently from any covariate. (A2) also includes the case of stratified randomization where the blocks are defined on some covariate X .*

2. *One important consequence of (A1) and (A2) is that for all purposes, we can now define the outcome variable Y without mentioning the assignment under which it is observed. More precisely, let $Y(d) = Y(0, d)\mathbf{1}_{\{T=0\}} + Y(1, d)\mathbf{1}_{\{T=1\}}$. For $d = 0, 1$, we have:*

$$\begin{aligned} \Pr(Y(d) \in A|X) &= \Pr(Y(0, d) \in A|T = 0, X) \Pr(T = 0|X) \\ &\quad + \Pr(Y(1, d) \in A|T = 1, X) \Pr(T = 1|X) \\ &= \Pr(Y(0, d) \in A|X) \Pr(T = 0|X) \\ &\quad + \Pr(Y(1, d) \in A|X) \Pr(T = 1|X), \quad (\text{by (A2)}) \\ &= \Pr(Y(0, d) \in A|X) = \Pr(Y(1, d) \in A|X), \quad (\text{by (A1)}) . \end{aligned}$$

For the rest of the paper we always assume (A1) and (A2) and write $Y_i(d)$ for the outcome of unit i that receives treatment d without necessarily mentioning its assignment.

The most important and certainly most controversial assumptions in randomized experiments are the next two assumptions.

Assumption **(A3)**:

$$(Y_i(0), Y_i(1)) \perp C_i|(T_i, X_i), \quad 1 \leq i \leq n, \quad (5)$$

Define $e(x) = \Pr(D = 1|X = x)$ as the probability of taking Treatment 1. This function is called the propensity score of the experiment.

Assumption **(A4)**: For any $x \in X$,

$$0 < e(x) < 1. \quad (6)$$

Assumption (A3) asserts that noncompliance is ignorable given the covariate vector X so that it can affect the outcome $Y(i)$ only through X , and (A4) is the so-called overlapping assumption that says that every unit have a positive chance of taking both treatments.

These assumptions are well-known and are at the heart of most covariate-based adjustments that have been proposed in observational studies (see e.g. Rosenbaum (2002)). In practice, they are very difficult to check. Most of the time one has to resort to specific knowledge of the phenomenon under study to judge whether these assumptions are reasonable given the covariates at hand. Our objective in this work is to show, in some special cases where these assumptions are known not to hold, that it is still possible to limit the bias in estimating the causal effect of the treatment.

Assumption (A4) also allows us to define $\mathbb{E}(Y(d)|X, D = d)$ as

$$\mathbb{E}[Y(d)|X, D = d] := \frac{\mathbb{E}[\mathbf{1}_{\{D=d\}}Y(d)|X]}{\Pr[D = d|X]}.$$

It can be easily seen that under the assumptions enumerated above, the receipt of treatment is ignorable and that $\tau_{\mathcal{P}}$ is consistently estimable.

Proposition 1 *Under (A1)-(A4), we have:*

$$Y(d) \perp D|X, \quad d \in \{0, 1\}. \tag{7}$$

This implies that $\mathbb{E}(Y(d)|X, D = d) = \mathbb{E}(Y(d)|X)$ so that $\tau_{\mathcal{P}}$ is consistently estimable from the data available.

Although Proposition 1 is directly usable to estimate $\tau_{\mathcal{P}}$, very often in practice, particularly when the dimension of the covariates X is large, it is more efficient to adjust for noncompliance on low-dimensional covariates. Rosembaum and Rubin (1983) introduced the concept of propensity scores and showed that it has that desired property. Their result in our context is as follows:

Proposition 2 *Assume (A1)-(A4). Let $b : \mathcal{X} \rightarrow \mathcal{Y}$ be a measurable function such that there exist measurable functions $g : \mathcal{Y} \rightarrow \mathbb{R}$ for which $e(x) = g(b(x))$, $x \in \mathcal{X}$. We have:*

(i)

$$X \perp D|b(X). \tag{8}$$

(ii)

$$(Y(0), Y(1)) \perp D|b(X), \tag{9}$$

Moreover if $b : \mathcal{X} \rightarrow \mathcal{Y}$ is any measurable function that satisfies (i) and (ii) above then necessarily, $e(x) = q(b(x))$ for some measurable function q .

2.2 Some practical implications of Propositions 1 and 2

Here we explore briefly some practical implications of Propositions 1 and 2 in estimating $\tau_{\mathcal{P}}$ in presence of noncompliance. There are many standard ways to adjust for covariates in randomized experiments that can be used here. We consider the use of the propensity scores and the regression functions. For more details, and for a review see Imbens (2003).

2.2.1 Weighting on propensity scores

It is clear from Proposition 1 that:

$$\begin{aligned}\mathbb{E}[Y(d)|X] &= \mathbb{E}[Y(d)|X, D = d] \\ &= \mathbb{E}\left[\frac{\mathbf{1}_{\{D=d\}}Y(d)}{\Pr(D = d|X)}|X\right].\end{aligned}$$

Therefore $\tau(X) = E\left\{\frac{\mathbf{1}_{\{D=1\}}Y(1)}{e(X)}|X\right\} - E\left\{\frac{\mathbf{1}_{\{D=0\}}Y(0)}{1-e(X)}|X\right\}$. If $\hat{e}(x)$ denotes a consistent estimate of e , $\tau_{\mathcal{P}}$ can be consistently estimated by:

$$\hat{\tau}_1 = \frac{1}{n} \sum_{i=1}^n \left(\frac{\mathbf{1}_{\{D_i=1\}}}{\hat{e}(X_i)} - \frac{\mathbf{1}_{\{D_i=0\}}}{1 - \hat{e}(X_i)} \right) Y_i, \quad (10)$$

where $Y_i = \mathbf{1}_{\{D_i=1\}}Y_i(1) + \mathbf{1}_{\{D_i=0\}}Y_i(0)$ is the outcome observed on the i -th unit.

We can estimate $e(x)$ with various parametric or nonparametric methods using the observed value of the variable (D, X) . In the examples that we consider below, we use a parametric logistic regression model.

The estimator (10) is well known in randomized experiments. It has been shown by Hirano et al (2003) (Theorem 1) that under some regularity conditions, if e_1 and e_0 are consistently estimated, $\hat{\tau}_1$ is consistent, efficient and asymptotically normal.

2.2.2 Correction based on the regression function

Let $\mu_d(x) = E(Y(d)|X)$. Again from Proposition 1 we have $E(Y(d)|X, \mathbf{1}_{\{D=d\}}) = E(Y(d)|X)$. Therefore, we can use the observed values on (Y, X) on the units that received Treatment d to obtain a consistent estimate $\hat{\mu}_d(x)$ for $\mu_d(x)$, $d = 0, 1$. This can be done using various parametric and nonparametric regression methods. Then a consistent estimate for τ is given by:

$$\hat{\tau}_2 = \frac{1}{n} \sum_{i=1}^n (\hat{\mu}_1(X_i) - \hat{\mu}_0(X_i)). \quad (11)$$

2.2.3 Subclassification and matching on the propensity scores

As permitted by Proposition 2, the propensity score can also be used to generate a new estimator that compares the outcome variables for similar levels of the propensity score, which is the so-called sub-classification on the propensity score method. Suppose we have an estimate \hat{e} of the propensity score e . Let $0 = b_1 < b_2 \cdots < b_{L+1} = 1$ be a subdivision of $[0, 1]$ constructed such that the values $\hat{e}(X_i)$ for which $\hat{e}(X_i) \in B_l := [b_l, b_{l+1}]$ are approximately equal. For $l = 1, \dots, L$, let:

$$n_l = \sum_{i=1}^n \mathbf{1}_{B_l}(\hat{e}(X_i)), \quad (12)$$

be the size of B_l , and

$$\hat{\tau}^{(l)} = \frac{\sum_{i=1}^n Y_i D_i \mathbf{1}_{B_l}(\hat{e}(X_i))}{\sum_{i=1}^n D_i \mathbf{1}_{B_l}(\hat{e}(X_i))} - \frac{\sum_{i=1}^n Y_i (1 - D_i) \mathbf{1}_{B_l}(\hat{e}(X_i))}{\sum_{i=1}^n (1 - D_i) \mathbf{1}_{B_l}(\hat{e}(X_i))}. \quad (13)$$

be the estimate of the causal effect for units in B_l . Then the subclassification estimator for $\tau_{\mathcal{P}}$ is given by:

$$\hat{\tau}_3 = \sum_{l=1}^L \frac{n_l}{n} \hat{\tau}^{(l)}. \quad (14)$$

The matching method (see e.g. Imbens (2003) and Diamond and Seckon (2005)) is a special case of subclassification with one treated unit and a fixed number M of control units (typically $M = 1$) per B_l . Matching is particularly useful when the number of control units is much larger than the number of treated units.

3 Inference under violated assumptions

3.1 Effect of unobserved covariates

In this section, we no longer assume (A3). So it is possible that there is some unobserved covariate that is linked to the outcome variable and to the compliance behaviour. As discussed in the introduction, it is particularly difficult to make valid statistical inference in randomized experiments with noncompliance when the noncompliance behaviour is not ignorable. That is when Assumption (A3) doesn't hold. When faced with nonignorable noncompliance behaviour, the general approach in practice seems to include as much covariables as possible in the analysis. But as shown by Spirtes (1997) when important covariables are unobserved, it is perfectly possible that controlling for all the observed covariates increases the estimation bias. In other words, unless all the important covariates related to the experiment are observed, adjusting for the observed covariates can actually increase the estimation bias.

In this work, we argue that in randomized experiment with noncompliance, one should be more caution and select only the covariates that can reproduce roughly well the selection bias of the experiment. We start with a theoretical result that will help clarify what we meant in the sentence above. Then we will discuss the details of how to implement the approach.

Formally, we want to answer two questions. What is the bias incurred if we proceed in estimating the causal effect τ_p without the unobserved covariate? Can we expect to reduce that bias compared to the situation where no covariates are controlled for? To simplify the notation, we write $e_d(x) = \Pr(D = d|X = x)$, $d = 0, 1$, so that $e_1(x) = e(x)$ and $e_0(x) = 1 - e(x)$. Also recall that $\mu_d(X) = \mathbb{E}(Y(d)|X)$.

Theorem 3 *Assume (A1), (A2), (A4) and for $d = 0, 1$, define*

$$\varepsilon_d(X) = \mathbb{E} [\mathbf{1}_{\{D=d\}}Y(d)|X] - e_d(X)\mu_d(X). \quad (15)$$

Then for $d = 0, 1$,

$$\mathbb{E} [Y(d)|X, \mathbf{1}_{\{D=d\}}] = \mathbb{E} \left[\frac{\mathbf{1}_{\{D=d\}}Y(d)}{e_d(X)} | X \right] = \mathbb{E}(Y(d)|X) + \frac{\varepsilon_d(X)}{e_d(X)}. \quad (16)$$

A first approximation of the expectation of the bias term $B_d(X) := \frac{\varepsilon_d(X)}{e_d(X)}$ gives:

$$\mathbb{E} [B_d(X)] \approx \delta_d [\text{Cov}(\mathbf{1}_{\{D=d\}}, Y(d)) - \text{Cov}(e_d(X), \mu_d(X))], \quad (17)$$

where $\delta_d = \Pr(D = d)^{-1} \left[\frac{\mathbb{E}(e_d^2(X))}{[\mathbb{E}(e_d(X))]^2} - \text{Cov} \left[\frac{e_d(X)}{\mathbb{E}(e_d(X))}, \frac{\varepsilon_d(X)}{\mathbb{E}(\varepsilon_d(X))} \right] \right]$.

Proof. Equation (16) is trivial. For random variables X and Y with a finite second moment, it is easily seen that:

$$\frac{X}{Y} - \frac{\mathbb{E}(X)}{\mathbb{E}(Y)} = \frac{X\mathbb{E}(Y) - Y\mathbb{E}(X)}{(\mathbb{E}(Y))^2 \left[1 + \frac{Y - \mathbb{E}(Y)}{\mathbb{E}(Y)} \right]} \approx \frac{X\mathbb{E}(Y) - Y\mathbb{E}(X)}{(\mathbb{E}(Y))^2} \left(1 - \frac{Y - \mathbb{E}(Y)}{\mathbb{E}(Y)} \right)$$

Taking the expectation yields:

$$\mathbb{E} \left[\frac{X}{Y} \right] \approx \frac{\mathbb{E}(X)}{\mathbb{E}(Y)} \left(\frac{\mathbb{E}(Y^2)}{(\mathbb{E}(Y))^2} - \frac{\text{Cov}(X, Y)}{\mathbb{E}(X)\mathbb{E}(Y)} \right)$$

Since $E(e_d(X)) = \Pr(D = d)$ and $E(\varepsilon_d(X)) = \text{Cov}(\mathbf{1}_{\{D=d\}}, Y(d)) - \text{Cov}(e_d(X), \mu_d(X))$, applying the approximation formula above with $X = \varepsilon_d(X)$ and $Y = e_d(X)$ gives:

$$\mathbb{E} \left[\frac{\varepsilon_d(X)}{e_d(X)} \right] \approx \delta_d [\text{Cov}(\mathbf{1}_{\{D=d\}}, Y(d)) - \text{Cov}(e_d(X), \mu_d(X))], \quad (18)$$

where δ_d is given as above. ■

Since (A3) doesn't hold, if we use the data from the experiment to estimate $\mu_d(X) = \mathbb{E}(Y_d|X)$, what we will actually be estimating is $\mathbb{E}(Y_d|X, \mathbf{1}_{\{D=d\}})$. The bias incurred is $\mathbb{E}(Y_d|X, \mathbf{1}_{\{D=d\}}) - \mu_d(X)$ and Theorem 3 shows that the expected bias is approximately $\mathbb{E}[B_d(X)] \approx \delta_d [Cov(\mathbf{1}_{\{D=d\}}, Y(d)) - Cov(e_d(X), \mu_d(X))]$. Note that $Cov(\mathbf{1}_{\{D=d\}}, Y(d)) / \Pr(D = d)$ is the overall selection bias due to noncompliance in estimating $\mathbb{E}(Y(d))$. Now, the term $Cov(e_d(X), \mu_d(X)) / \Pr(D = d)$ can be seen as the estimate provided by the covariate X of that selection bias $Cov(\mathbf{1}_{\{D=d\}}, Y(d)) / \Pr(D = d)$. Therefore if X provides a correct estimate of the selection bias, that is if $Cov(e_d(X), \mu_d(X)) \approx Cov(\mathbf{1}_{\{D=d\}}, Y(d))$ for $d = 0, 1$, then adjusting on X will reduce the selection bias. But if X is such that $Cov(e_d(X), \mu_d(X))$ and $Cov(\mathbf{1}_{\{D=d\}}, Y(d))$ are very different quantities, it is best not to adjust on X as this will increase the estimation bias. When prior information exists, this result can be used to screen to set of observed covariates in order to used the most relevant.

3.1.1 Covariates selection with prior information

As discussed above, our selection method consists in finding all the covariates X for which $Cov(e_d(X), \mu_d(X))$ and $Cov(\mathbf{1}_{\{D=d\}}, Y(d))$ are roughly equal. Since neither of these quantities can be computed is most experiment, we settle with a less rigorous rule: select all the covariates for which $Cov(e_d(X), \mu_d(X))$ and $Cov(\mathbf{1}_{\{D=d\}}, Y(d))$ have the same signs. Although the actual values cannot be estimated, in many cases, it is possible to use prior information on the experiment to find an estimate of the signs of the quantities.

Since $Cov(\mathbf{1}_{\{D=d\}}, Y(d)) / \Pr(D = d) = E(Y(d)|D = d) - E(Y(d))$, we need to compare the overall expected outcome to the expected outcome of units that choose to receive d . Such information is available in many studies.

To obtain the sign of $Cov(e_d(X), \mu_d(X))$, we rely on the following well known result. Let Y be a random variable and f and g two monoton functions. If f and g has the same monotonicity then $Cov(f(Y), g(Y)) \geq 0$ otherwise $Cov(f(Y), g(Y)) \leq 0$. In our case, the monotonicity of $e_d(x)$ can be obtained by estimating that function from data. The monotonicity of $\mu_d(x) = \mathbb{E}(Y(d)|X = x)$ cannot be estimated from the data in general. But in many studies this information can be available.

Here is an example of how to apply the method. Assume that from prior information, we know the sign of $E(Y(d)|D = d) - E(Y(d))$, $d = 0, 1$. For example, assume that $E(Y(1)|D = 1) - E(Y(1)) \geq 0$ and that $E(Y(0)|D = 0) - E(Y(0)) \leq 0$. Note that this information does not say anything about the effect of the treatment but is related instead to what is known about the noncompliance behavior. For example, $E(Y(1)|D = 1) - E(Y(1)) \geq 0$ simply says, speaking about

the outcome when Treatment 1 is received, that those who choose to receive that treatment happen to have (on average) a larger value of the outcome. Let X be the observed covariate vector and also assume that prior information exists which gives us the monotonicity of $\mu_d(x)$. To fix the ideas, assume that $\mu_d(x)$ is nondecreasing for $d = 0, 1$. Let $\hat{e}_1(x)$ be an estimate of $e_1(x)$ obtained from the data.

Our rule to decide whether to control for X is as follows. If we find out that $\hat{e}_1(x)$ is nondecreasing in x , then it is likely that $Cov(e_1(X), \mu_1(X)) \geq 0$ which is consistent with $E(Y(1)|D=1) - E(Y(1)) \geq 0$. Similarly, since $\mu_0(x)$ is nondecreasing, we likely have $Cov(e_0(X), \mu_0(X)) \leq 0$ which is consistent with $E(Y(0)|D=0) - E(Y(0)) \leq 0$. Therefore X will be included in our list of potentially useful covariates. But if we find out that $\hat{e}_1(x)$ is nonincreasing (in x) then X is likely to increase the estimation bias and should not be used.

The method is repeated for each measured covariate to determine the final set of covariates to be used to control for the noncompliance bias. We use that method to re-analyze a randomized experiment with encouragement designed to measure the effect of an influenza vaccine. See Section 4.2.

3.2 Inference in taboo population

In this section, we deal with situations where it is not actually possible to apply the treatments to the whole population. Assume that there exist thresholds $-\infty < z^{(t)} \leq \infty$ and continuous functions $Z^{(t)} : X \rightarrow R$ ($t = 0, 1$) such that any unit i with covariate X_i for which $Z^{(t)}(X_i) > z^{(j)}$ cannot be given Treatment t (or equivalently that such unit will never comply with Treatment t). Therefore we have $e_1(x) = e(x) = 0$ for $Z^{(1)}(x) > z^{(1)}$ and $e_0(x) = 1 - e(x) = 0$ for $Z^{(0)}(x) > z^{(0)}$ so that (A4) does not hold. We call the sub-populations $\{x : Z^{(1)}(x) > z^{(1)}\}$ and $\{x : Z^{(0)}(x) > z^{(0)}\}$ taboo sub-populations. A common example where this model arises naturally is in medical studies where units are first screened on some pre-treatment variables, the so-called eligibility criteria, in order to determine their participation to the experiment. Although only a subset of the whole population can participate to the experiment, sometimes one is still interested in inferring the causal effect of the treatment on the whole population from such an experiment. We argue here that if a global model for the treatment effect is plausible then the effect of the treatment over the whole population is estimable. The argument is very simple and is based on the following seen conditional expectation:

$$\mathbb{E} \left[Y(d) | X, \mathbf{1}_{\{Z^{(d)}(X) \leq z^{(d)}\}} \right] = \mathbf{1}_{\{Z^{(d)}(X) \leq z^{(d)}\}} \mathbb{E}[Y(d) | X]. \quad (19)$$

From this, it follows that if we have an iid sample $(Y_i(d), X_i)$ where $Z^{(d)}(X_i) \leq z^{(d)}$, using an appropriate regression model, we can obtain a consistent estimate for $\mathbb{E}[Y(d) | X = x]$

on $\{x : Z^{(d)}(x) \leq z^{(d)}\}$. If $\mathbb{E}[Y(d)|X = x]$ has a global parametric form, then the regression analysis will consistently estimate those parameters and $\mathbb{E}[Y(d)|X = x]$ can then be estimated outside $\{x : Z^{(d)}(x) \leq z^{(d)}\}$. That is the basic idea that we are proposing. In practice, because of the sampling scheme (the way the units are screened and assigned to the treatments), the final sample $(Y_i(d), X_i)$ where $Z^{(d)}(X_i) \leq z^{(d)}$ is generally not iid. There are many ways to design the sampling of the experiment. We mention two basic methods. In one form a fixed number of units is screened and randomized to the two arms of the experiments and the numbers of units by treatment is finally random. In the other form, units are screened and randomized to the treatments until a predetermined number of units are obtained for both treatments. Here the total number of units screened is random. We assume the latter sampling method. That is, we assume that we have an infinite reservoir of units that are screened and assigned until the first time a given number of units, say n_1 , is assigned to Treatment 1.

For $i \geq 1$, let $(Y_i^*(0), Y_i^*(1), X_i^*, T_i^*)$ be a sequence of i.i.d. random variables such that T_i^* and $(Y_i^*(0), Y_i^*(1))$ are independent given X_i^* .

For $t \in \{0, 1\}$, define $\tau_0^{(t)} = 0$ and for $k \geq 1$ define

$$\tau_k^{(t)} := \inf \left\{ n > \tau_{k-1}^{(t)} : T_n^* = t, Z^{(t)}(X_n^*) \leq z^{(t)} \right\}. \quad (20)$$

We assume that $\rho_t := \Pr(T_1^* = t, Z^{(t)}(X_1^*) \leq z^{(t)}) > 0$ so that all these random times $\tau_k^{(t)}$ are finite.

The random time $\tau_1^{(0)}$ is the first unit assigned to Treatment 0 that is under the threshold $z^{(0)}$ and thus will receive that treatment. Similarly, $\tau_k^{(0)}$ is the k -th unit assigned to Treatment 0 that satisfies the constraint and thus will receive that treatment. Implicitly, we assume that there is no further noncompliance issue after a unit satisfies the constraint. Similar explanations hold for the stopping times $\tau_k^{(1)}$. The sampling is stopped after n_1 units have been assigned to Treatment 1. Let m be the number of units screened before obtaining the n_1 units assigned to Treatment 1. Let n_0 be the number of unit assigned to Treatment 0 during the sampling. Here m and n_0 are random. m follows a Negative Binomial distribution with parameters n_1 and ρ_1 and given m , n_0 follows a Binomial distribution with parameters m and ρ_0 . Recall that for $t = 0, 1$, $\rho_t := \Pr(T_1^* = t, Z^{(t)}(X_1^*) \leq z^{(t)})$. Finally, we assume that all the covariates $(X_i^*)_{1 \leq i \leq m}$ tested during the sampling are kept in order to compute the causal effect for the entire population.

The idea we are about to use is simple. Although the special sampling used to obtain the two samples induces a dependence between the observations, it actually does not change the dependence between the outcome and the covariates of any one unit and our arguments above should work as well. To continue, we assume the following global representation of the conditional expectation of $Y^*(d)$ as functions of the covariates:

Assumption **(A5)**: There exist known measurable functions $(h_1^{(d)}, \dots, h_{p_d}^{(d)})$ and parameters $\beta^{(d)} = (\beta_1^{(d)}, \dots, \beta_{p_d}^{(d)})'$ such that

$$\mathbb{E}(Y^*(d)|X^* = x) = H_d(\beta, x) = \sum_{i=1}^{p_d} \beta_i^{(d)} h_i^{(d)}(x), \quad d = 0, 1. \quad (21)$$

The additive model postulated in Assumption (A5) is quite flexible. By allowing the user to specify the component functions $h_i^{(d)}$, that approach makes it easier to use prior knowledge available on the dependence between the outcome and the covariates. When no such information exists, It is always possible to use a more systematic set of functions like $1, X_1, X_2, \dots, X_p, X_1X_2, X_1X_3 \dots$, where here, X_i is the i -th component of X . Another possible approach is to use a nonparametric additive model. In this model, the regression function is taken as $H_d(\beta, x) = \sum_{i=1}^{p_d} h_i^{(d)}(x)$, where $h_i^{(d)}$ is an unknown function and estimated nonparametrically, for example as in Horowitz and Mammen (2004). The important point is that (A5) is a working assumption and other (more general) global modelling is also possible.

Let $\hat{\beta}^{(1)}$ be the Ordinary Least Square (OLS) estimate of $\beta^{(1)}$ in the regression of the outcome Y on $(h_1^{(1)}(X), \dots, h_{p_1}^{(1)}(X))$ obtained from the sample $(Y_{\tau_i^{(1)}}, X_{\tau_i^{(1)}})_{1 \leq i \leq n_1}$. Similarly, define by $\hat{\beta}^{(0)}$ the OLS estimate of $\beta^{(0)}$ computed from the sample $(Y_{\tau_i^{(0)}}, X_{\tau_i^{(0)}})_{1 \leq i \leq n_0}$. We introduce the estimator

$$\hat{\tau}_4 := \frac{1}{m} \sum_{i=1}^m \left[H(\hat{\beta}^{(1)}, X_i^*) - H(\hat{\beta}^{(0)}, X_i^*) \right]. \quad (22)$$

It turns out that:

Theorem 4 *Assume (A1)-(A3) and (A5). Then, $\mathbb{E}(\hat{\tau}_4) \rightarrow \tau_p$ as $n_1 \rightarrow \infty$, where τ_p is the treatment effect over the whole population.*

Proof. It suffices to show that $\mathbb{E}\left(\frac{1}{m} \sum_{i=1}^m H(\hat{\beta}^{(1)}, X_i^*)\right) \rightarrow \mathbb{E}(Y(1))$ as $n_1 \rightarrow \infty$. A similar argument will apply for $\mathbb{E}\left(\frac{1}{m} \sum_{i=1}^m H(\hat{\beta}^{(0)}, X_i^*)\right)$ and the Theorem will be proved.

It is not hard to see that $\mathbb{E}\left(Y_{\tau_i^*}^* | m, X_1^*, \dots, X_m^*\right) = \beta^{(1)'} h^{(1)}(X_{\tau_i^*}^*)$, where $h^{(1)}(x) = (h_1^{(1)}(x), \dots, h_{p_1}^{(1)}(x))'$. Therefore:

$$\mathbb{E}\left(\frac{1}{m} \sum_{i=1}^m H(\hat{\beta}^{(1)}, X_i^*)\right) = \beta^{(1)'} \mathbb{E}\left(\frac{1}{m} \sum_{i=1}^m h^{(1)}(X_i^*)\right). \quad (23)$$

The proof will be finished once we prove that for any $1 \leq j \leq p_1$, $\mathbb{E}\left(\frac{1}{m} \sum_{i=1}^m h_j^{(1)}(X_i^*)\right) \rightarrow \mathbb{E}\left(h_j^{(1)}(X_1^*)\right)$.

If i is such that $T_i = 1$, and $Z^{(1)}(X_i^*) \leq z^{(1)}$, then:

$$\mathbb{E} \left(h_j^{(1)}(X_i^*) | m \right) = \frac{1}{\rho_1} \int_{D_1} h_j^{(1)}(x) l_1(x) \mu_X(dx),$$

where μ_X is the distribution of X , $D_1 = \{x \in \mathcal{X} : Z^{(1)}(x) \leq z^{(1)}\}$, $l_1(x) = \Pr(T^* = 1 | X^* = x)$.

If i is such that $T_i = 0$, and $Z^{(0)}(X_i^*) \leq z^{(0)}$, then:

$$\mathbb{E} \left(h_j^{(1)}(X_i^*) | m \right) = \frac{1}{\rho_0} \int_{D_0} h_j^{(1)}(x) l_0(x) \mu_X(dx),$$

where $D_0 = \{x \in \mathcal{X} : Z^{(0)}(x) \leq z^{(0)}\}$, $l_0(x) = 1 - l_1(x)$.

And if i is such that ($T_i = 1$, and $Z^{(1)}(X_i^*) > z^{(1)}$) or ($T_i = 0$, and $Z^{(0)}(X_i^*) > z^{(0)}$), then:

$$\mathbb{E} \left(h_j^{(1)}(X_i^*) | m \right) = \frac{1}{1 - \rho_0 - \rho_1} \left(\int_{\bar{D}_1} h_j^{(1)}(x) l_1(x) \mu_X(dx) + \int_{\bar{D}_0} h_j^{(1)}(x) l_0(x) \mu_X(dx) \right),$$

where \bar{A} represents the complement of A .

From these expressions, we deduce that:

$$\begin{aligned} \mathbb{E} \left(\frac{1}{m} \sum_{i=1}^m h_j^{(1)}(X_i^*) \right) &= \frac{1}{\rho_1} \mathbb{E} \left(\frac{n_1}{m} \right) \int_{D_1} h_j^{(1)}(x) l_1(x) \mu_X(dx) \\ &\quad + \frac{1}{\rho_0} \mathbb{E} \left(\frac{n_0}{m} \right) \int_{D_0} h_j^{(1)}(x) l_0(x) \mu_X(dx) \\ &\quad + \frac{1}{1 - \rho_0 - \rho_1} \left(1 - \mathbb{E} \left(\frac{n_0}{m} \right) - \mathbb{E} \left(\frac{n_1}{m} \right) \right) \left[\int_{\bar{D}_1} h_j^{(1)}(x) l_1(x) \mu_X(dx) \right. \\ &\quad \left. + \int_{\bar{D}_0} h_j^{(1)}(x) l_0(x) \mu_X(dx) \right]. \end{aligned}$$

Given m n_0 distributed as a Binomial distribution with parameters m, ρ_0 , therefore $\frac{1}{\rho_0} \mathbb{E} \left(\frac{n_0}{m} \right) = 1$.

As we saw, m is distributed as a Negative Binomial distribution with parameters n_1 and ρ_1 , therefore

$\frac{1}{\rho_1} \mathbb{E} \left(\frac{n_1}{m} \right) \rightarrow 1$ as $n_1 \rightarrow \infty$. And we get:

$$\mathbb{E} \left(\frac{1}{m} \sum_{i=1}^m h_j^{(1)}(X_i^*) \right) \longrightarrow \int h_j^{(1)}(x) \mu_X(dx), \quad \text{as } n_1 \rightarrow \infty$$

after some simplifications. ■

4 Application

4.1 A simulation example

Here we conduct a Monte Carlo experiment to illustrate that Intent-to-Treat and the Local Average Treatment Effect can be seriously biased as estimates of the treatment effect over the

population. Let X_1, X_2 be two covariates defined on some population and assume that $X_1 \sim N(60, 15)$ and $X_2 \sim N(2, 0.5)$. We would like to estimate the causal effect of a treatment (Treatment 1) compared to Treatment 0 over that population. We take $2N$ (with $N = 850$) units and randomly evenly split them over the two treatments such that for any unit i , $P(T_i = 1) = 1/2$. We assume that there is full compliance with assignment to Treatment 0; that is, $D_i = T_i$ if $T_i = 0$. But compliance to Treatment 1 is not perfect. More specifically, we have

$$P(C_i = 1 | T_i = 1, X_1 = x_1, X_2 = x_2) = \frac{\exp\{\alpha_0 + \alpha_1 x_1 + \alpha_2 x_2\}}{1 + \exp\{\alpha_0 + \alpha_1 x_1 + \alpha_2 x_2\}},$$

with $(\alpha_0, \alpha_1, \alpha_2) = (5, -.005, -1)$.

Let Y be the response or outcome variable. We make the exclusion-restriction assumption and assume that if a unit receives Treatment j , $j = 0, 1$, we observe $Y(i)$ and

$$Y(i) | (X_1 = x_1, X_2 = x_2) \sim N(\beta_{i,0} + \beta_{i,1}x_1 + \beta_{i,2}x_2 + \beta_{i,3}x_1x_2, 1),$$

where $(\beta_{0,0}, \beta_{0,1}, \beta_{0,2}, \beta_{0,3}) = (-3.8, 0.01, 1, 0.01)$ and $(\beta_{1,0}, \beta_{1,1}, \beta_{1,2}, \beta_{1,3}) = (5, -0.05, -1, 0)$. The parameters are chosen such that $E(Y(0)) = 0$ and $E(Y(1)) = 0$ and therefore the treatment has no effect on the population. As the reader can see, the example is constructed in such a way that units with relatively large values of (X_1, X_2) tend not to comply to assignment to Treatment 1 which appears to be “harmful ” for that sub-population. They switch to Treatment 0 which is more beneficial for them. We compare the ITT estimate and the LATE estimate with $\hat{\tau}_1$, $\hat{\tau}_2$ and $\hat{\tau}_3$. Let Y_i be the response observed on the i -th unit. We compute the ITT and LATE estimates as:

$$ITT = \frac{1}{N} \left(\sum_{i=1}^{2N} Y_i T_i - \sum_{i=1}^{2N} Y_i (1 - T_i) \right),$$

$$LATE = \frac{ITT}{\frac{1}{N} \left(\sum_{i=1}^{2N} D_i T_i - \sum_{i=1}^{2N} D_i (1 - T_i) \right)}.$$

We have simulated this model 250 times to compute the means of the various estimates as well as their standard errors. Table 1 shows the results.

Table 1: Means and standard errors of the causal effect estimates.

	ITT	LATE	$\hat{\tau}_1$	$\hat{\tau}_2$	$\hat{\tau}_3$
Mean	0.3323	0.6643	0.0402	0.0005	0.0286
Std error	0.0688	0.1379	0.0778	0.0756	0.0565

[0pt] Graph 1: Density estimates for the distribution of the 5 estimators used.

(insert Graph 1 here)

It is clear from this example that Intent-to-Treat and the Local Average Treatment Effect are not good estimates of the real treatment over the population. When noncompliance can be fairly well-predicted from the covariates available, as in this example, the estimates proposed in this paper provide better estimates. In case of LATE, that estimate is correct if we restrict our interest to the sub-population of compliers. But as discussed in the introduction, such a sub-population is not well-defined here; thus, the estimate doesn't tell us anything about the real effect of the treatment over the whole population.

4.2 Assessing the effect of an influenza vaccine

We reanalyze a randomized experiment conducted by McDonald et al (1992) to assess the efficacy of influenza vaccination in reducing flu-related morbidity in high-risk populations. The study has also been reanalyzed by Hirano et al (2000). The study was designed to circumvent the ethical problem of denying the vaccine to part of the population while giving it to others as in a classical randomized experiment. Physicians in a public hospital were randomly assigned either to the treatment group or to the control group. When a high-risk patient whose physician is in the treatment group is scheduled for an appointment, a computer-generated letter is sent to the physician to encourage her to vaccinate the patient. According to the US Public Health Service Criteria, patients over 65 years of age or with chronic lung disease, asthma, diabetes mellitus, congestive heart failure or severe renal or hepatic failure are at high-risk of flu.

For a given patient i , we observe T_i , which is 1 if that person's physician is in the treatment group and a letter is sent to encourage the physician to vaccinate the patient, or 0 otherwise. The patient may choose to comply with its assignment T_i in which case $C_i = 1$ or not, in which case $C_i = 0$. We also observe the variable D_i which is 1 if the patient actually receives the flu shot and 0 otherwise. The outcome variable is a binary variable $Y_i(t, d)$ which is 1 if the patient i is subsequently admitted in hospital for flu during the next winter, or 0 otherwise. The age (*age*) and various medical condition variables are also observed on the patients before the assignment. These are *copd* $\in \{0, 1\}$ for chronic pulmonary disease, 1 is 'yes'; *dm* $\in \{0, 1\}$ for diabetes mellitus, 1 is 'yes'; *heartd* $\in \{0, 1\}$ for heart disease, 1 is 'yes'; *renal* $\in \{0, 1\}$ for severe renal failure, 1 for 'yes'; *liverd* $\in \{0, 1\}$ for chronic hepatic failure, 1 is 'yes'. The gender *sex* $\in \{0, 1\}$, 1 for 'male' and *race* $\in \{0, 1\}$, 1 for 'Caucasian', 0 for 'non-caucasian' are also given. We ignore these last two variables in our analysis. See table 2, for a summary of the data.

[0pt] Table 2: Sample means for the variables of interest over the subpopulations defined by Z and D .

	Z	D	Y	age	copd	dm	heartd	liverd	sex	race
	0.5145	0.2502	0.0853	65.27	0.2820	0.2786	0.5736	0.0031	0.6683	0.6550
Z=0	0	0.1893	0.0929	65	0.2894	0.2829	0.5644	0.0036	0.6523	0.6487
Z=1	1	0.3077	0.0781	65.5	0.2751	0.2744	0.5822	0.0027	0.6834	0.6610
D=0	0.4750	0	0.0853	64.72	0.2625	0.2765	0.5660	0.0033	0.6793	0.6587
D=1	0.6327	1	0.0852	66.9	0.3408	0.2849	0.5964	0.0028	0.6355	0.6439

Following Hirano et al (2000), we ignore the clustering of the units through the physicians. We do this to keep the analysis simple and also because we do not have sufficient information on the clustering. Therefore we can consider the assignment to the flu shot as completely randomized so that (A2) holds. In contrast to Hirano et al (2000), we make the exclusion-restriction assumption (Assumption (A1)) that there is no difference in the distribution of the outcome between the patients who had a different assignments but received the same treatment. Given that the study involves mainly seniors most of them with high risk to influenza, it seems likely that there are all well aware of the threat posed by the flu season and the basic ways to protect themselves. Therefore it is very unlikely that a discussion with their physician will make a systematic difference.²

Finally we assume (A3') and (A4) and use the method developed in Section 3.1. It is very plausible and we shall assume that if all the patients were given the flu shot, those who choose voluntarily to take the shot would probably have a higher hospitalization rate:

$$\mathbb{C}ov(\mathbf{1}_{\{D=1\}}, Y(1)) \geq 0. \quad (24)$$

Similarly, if all the patients were denied the shot, those who choose not to take the shot would probably have a lower hospitalization rate:

$$\mathbb{C}ov(\mathbf{1}_{\{D=0\}}, Y(0)) \leq 0. \quad (25)$$

In relation with the covariates, another reasonable hypothesis is that for any covariate vector $X \in \{age, copd, dm, heartd, renal, liverd\}$: $E(Y(0)|X)$ and $E(Y(1)|X)$ are roughly nondecreasing in X . For example it seems reasonable to assume that older patient will have higher hospitalization rate than younger patients. Similarly, patients suffering from any above-mentioned disease is likely to have a higher hospitalization date than patients not suffering from that disease.

If we agree with the prior information on the study, then we could use this information to screen the observed covariates to determine which ones agree and are therefore likely to reduce the noncompliance bias. Table 2 shows $\Pr(D = 1|X)$ varies with X for the different covariates. When $X = age$, we estimate the $\Pr(D = 1|X)$ using a logistic regression and report the correlation

²Hirano et al (2000) seems to suggest otherwise.

between $\hat{\Pr}(D = 1|X)$ and X . When X is any one of the other binary variable we report the sample estimate of $\Pr(D = 1|X = 1) - \Pr(D = 1|X = 0)$.

From this table, and given the method outlined in Section 3.1.1, we are lead to adjust for the covariates *age, copd, dm, heartd, renal*.

Table 3: Estimate of the monotonicity between $\Pr(D = 1|X)$ and X when X is the different covariates observed.

age	copd	dm	heartd	liverd
0.9969	0.0726	0.0079	0.023	-0.0281

Based on these assumptions and the final set of selected covariates, we use the estimator $\hat{\tau}_2$, to estimate the causal effect of the treatment over the population and also over different subset of high risk patients. The results are summarized in Table 3. We compute the treatment effect for the entire population and for different high risk sub-populations. The standard errors are estimated by bootstrapping. From these results, we conclude that there is no evidence that the vaccination has any particular effect on the hospitalization rate in general. But for individuals at high risk, our analysis strongly suggests that the vaccination has improved the hospitalization rate for those high risk groups even though the effects are not statistically significant.

Table 4: Overall treatment effect and treatment effect for some high risk groups

	overall	age > 65	copd	dm	heartd
Mean	-0.0178	-0.0060	-0.1222	-0.0751	-0.0962
s.d.	0.1822	0.1882	0.2862	0.2795	0.2037

4.3 Effect of the Progresa program

In this example we analyse the impact of the Mexican Progresa poverty program on school attendance for children between 13 and 15 years of age. For a detailed description of Progresa, we refer the reader to Schultz (2004). Progresa was a vast mutli-year program implemented by the Mexican government to reduce poverty. It has been designed as a randomized experiment. The government has first identify all the localities where poverty was of concern. 495 localities were found. A census was conducted in October 1997 to collect information on income, consumption, assets etc... on all the households in these localities. The data from the census was used to determine a poverty index. Only very poor households (with poverty index below a threshold) were elligible for

Progresa. So the taboo population here is the households with poverty index above that threshold. The households are not directly randomized. The localities are randomly assigned to receive the program or to serve as control and all the households in that localities are treated accordingly. Out of the 495 poor localities, 314 have been assigned to receive the Program, the remaining localities serving as control. The Progresa program itself is made up of three components. There is an educational grant to facilitate and encourage the education of children. The program also gives an improved health services to all the members of the participating household. And a monetary transfer is provided to help satisfy their basic needs.

We estimate the effect of the program on the school attendance of children between 13 and 15 years old, 1 year after the beginning of Progresa. The covariates used in our analysis are based on the 1997 preliminary survey. The experiment offers a direct estimate of the effect of the program on “poor” children (children in eligible household). It is certainly also of interest to estimate the effect that the program would have on “rich” children (children in non-eligible household).

Following the approach outlined above, we used a logistic regression model to estimate the effect of the program on school attendance in the group of “poor” children. For each of the two arms of the experiment, if we assume that the logistic model is well specified, then the estimated model also gives a consistent estimate for the regression of school attendance on the covariates for the whole population in that arm of the experiment. Therefore a causal effect of Progresa for “rich” children and for the whole population can be inferred. The results are presented in Table 5. It appears according to these estimates that the effect of the program for “rich” children would be comparable although slightly less important than what has been observed for “poor” children.

Table 5: Estimate of Progresa impact on Mexican population. The estimates for “Rich” and “Overall” are the effect, as given by our model, that we would have observed if Progresa was extended to this group.

	Poor	Rich	Overall
Mean	0.2907	0.2627	0.2812
s.d.	0.0121	0.0139	0.0122

5 Conclusion

In the paper we have considered randomized experiments where some of the classical assumptions do not hold. Firstly, we have considered randomized experiments with noncompliance when the noncompliance is not ignorable. We Instead using all available covariates, a solution

that may perform badly, we have proposed a covariate selection method that select the best subset of covariates to adjust on in order to minimize the estimation bias. Our method requires that some prior information on the experiment be available. In the second part of the paper, we have considered experiments where an eligibility criteria based on some observed covariate is required to enter the different arms of the experiment. Despite the fact that only a subset of the population can enter the experiment, we have shown that a nonparametric model can be used to estimate the regression function of the treatments' effect on the covariates. We have used this model to estimate the causal effect of the treatment over the whole population.

References

- [1] Angrist, J.D., Imbens, G.W. and Rubin, D. (1996). Identification of causal effects using instrumental variables (with discussion), *JASA* 91 444-155
- [2] Diamond, A. Seckhon, J.S. (2005). Genetic Matching for estimating causal effects: a new method of achieving balance in observational studies. Technical report, Department of Government, Harvard University
- [3] Fisher, R. (1935). *The design of experiments*. Boyd, London
- [4] Hirano, K., Imbens, G. and Ridder, G. (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica* 71 1161-1189
- [5] Hirano, K., Imbens, G.W., Rubin, D.b. and Zhou, X.-H. (2000). Assessing the effect of an influenza vaccine in an encouragement design. *Biostatistics* 1 69-88
- [6] Holland, P. (1986). Statistics of causal inference. *Journal of American Statistical Association* 81 945-970
- [7] Horowitz, J. L. and Mammen, E. Nonparametric estimation of an additive model with a link function. *Annals of Statistics*, 32, 2412-2443.
- [8] Imbens, G. W. (2003). Semiparametric estimation of average treatment effects under exogeneity: a review, Technical Report, Department of Economics, UC Berkeley
- [9] Imbens, G. W. and Rubin, D. (1997). Bayesian Inference for causal effects in randomized experiments with noncompliance. *Annals of Statistics* 25 305-327

- [10] Lee, Y. Ellenberg, J. Hirtz, D. and Nelson, K. (1991). Analysis of clinical trials by treatment actually received: is it really an option? *Statistics in medicine* 10 1595-1605
- [11] McDonald, C., Hiu, S., Tierney, W. (1992). Effects of computer reminders for influenza vaccination on morbidity during influenza epidemics *MD Computing* 9 304-312
- [12] Neyman J. (1923). On the application of probability theory to agricultural experiments. essay on principles. section 9 (with discussion) translated in *Statistical Sciences* vol 5, No 4 465-480
- [13] Rosembaum, P. (2002). *Observational Studies*. Springer, New York.
- [14] Rosembaum, P. R. and Rubin, D.(1983). The central role propensity score in observational studies for causal effects. *Biometrika* 76 41-55
- [15] Rubin, D. (1974). Estimating causal effects of treatments in randomized and non randomized studies. *Journal of Educational Psychology* 66 688-701
- [16] Rubin D. (1978). Bayesian Inference for causal effects: the role of randomization. *Annals of statistics* 6 34-58
- [17] Spirtes, P. (1997). Limits on causal inference for statistical data. presented at the American Economics Association meetings
- [18] Wantchekon, LO. (2003). Clientelism and voting and behavior: Evidence from a field experiment in Benin *World Politics* 55 399-422