CAE Working Paper #05-11

# Missing Treatments

by

Francesca Molinari

May 2005

# Missing Treatments*

Francesca Molinari†

Department of Economics, Cornell University

This version: May 2005

**Abstract**

The existing literature on treatment effects assumes perfect observability of the treatments received by the population of interest. Even in cases of imperfect compliance, it is usually assumed that both the assigned and administered treatment are observed (or missing completely at random). This paper abandons such assumptions. Imperfect observability of the received treatment can arise as a result of survey nonresponse in observational studies, or noncompliance with randomly assigned treatments that are not directly monitored. I study the problem in the context of observational studies. I derive sharp worst case bounds without assuming anything about treatment selection, and I show that the bounds are a function of the available prior information on the distribution of the missing treatments. Under the maintained assumption of monotone treatment response, I show that no prior information on the distribution of missing treatments is necessary to get sharp informative bounds. I apply the methodologies recently proposed by Imbens and Manski (2004) and Chernozhukov, Hong, and Tamer (2004) to derive two types of confidence intervals for the partially identified parameters. The results are illustrated with an empirical analysis of drug use and employment using data from the National Longitudinal Survey of Youth.

**Keywords:** Missing Data; Bounds; Partial Identification; Treatment Effect.

*This paper is a revised version of Chapter 1 of my Northwestern University Ph.D. Dissertation "Contaminated, Corrupted, and Missing Data" (2003). I am grateful to my dissertation committee: Tim Conley, Joel Horowitz, Chuck Manski and Rosa Matzkin. I am especially indebted with Chuck Manski for his constant encouragement and advice. I thank Joe Altonji, Gadi Barlevy, Levon Barseghyan, Arie Beresteanu, Juan Carlos Chavez-Martin del Campo, Jeff Dominitz, Luojia Hu, Daniel McMillen, Lyndon Moore, Joerg Stoye, Chris Taber, Tymon Tatur and Tim Vogelsang for helpful discussions. I benefitted from the opportunity to present this work at the 2002 Meeting of the Midwest Economic Association and at the 2002 North American Summer Meetings of the Econometric Society. All remaining errors are my own. This research was partially supported by a Graduate Research Grant from the University Research Grants Committee at Northwestern University.

†492 Uris Hall, Ithaca, NY 14853-7601, fm72@cornell.edu.

# 1 Introduction

Part of the existing literature on programme evaluation has examined what can be inferred about a treatment effect of interest when the observability of some of the relevant variables is imperfect, and when noncompliance to assigned treatments prevents identification in randomized experiments. Little (1992), Robins, Rotnizky, and Zhao (1994), and Wang, Wang, Zhao, and Ou (1997) formulate sets of assumptions strong enough to achieve point identification for the case of missing covariate data. Imbens and Pizer (1999) show that in randomized experiments with complete random assignment of treatment, the assumption of covariate data missing at random can be tested. They develop models that allow to achieve point identification, are consistent with the restrictions implied by the complete random assignment of treatment, and are as close to the assumption of data missing at random as possible. Horowitz and Manski (2000) study the problem of missing outcome and covariate data in randomized experiments. They derive sharp bounds on the distribution of outcomes conditional on covariates without invoking untestable assumptions on the missing data mechanism. Angrist, Imbens, and Rubin (1996) address the problem of imperfect compliance in classical randomized experiments; they pose a set of assumptions under which it is possible to identify the treatment effect within the subpopulation of persons who comply with the assigned treatment. Robins (1989) and Balke and Pearl (1994), (1995) and (1997) study similar identification problems under weaker assumptions; in particular Balke and Pearl (1997) make use of the statistical independence between the response functions and the assigned treatments to propose alternative, assumption-free sharp bounds for assessing the average effect of treatment over the population as a whole. A different problem is addressed by Hotz, Mullin, and Sanders (1997), who study what can be learned about treatment effects when one uses a contaminated instrumental variable, i.e. when a mean-independence assumption holds in a population of interest, but the observed population is a mixture of the population of interest and one in which the assumption doesn't hold.

A common feature in this literature is the assumption of perfect observability of the received treatment (both in randomized experiments and observational studies), or ignorability of the data with missing treatments. This assumption is often at odds with the empirical evidence. Survey nonresponse can affect the observability of a variable whose effect is under study, and often it is not plausible to assume that the decision to respond to a specific question in a survey is random. For example, researchers face relatively high nonresponse rates when studying the effect of "problem drinking" or drug abuse on labor market outcomes (see for example Mullahy and Sindelar (1995), Kaestner (1991), (1994) and (1998), and Kaestner and Grossman (1995)); at the same time, it doesn't seem plausible to assume that the fraction of drug abusers or alcoholics is identical for respondents and nonrespondents (see for example Pepper (2001)). Another example comes from

Robins (1997), who describes a nonrandomized experiment in which observation of a treatment group and a control group of women who had a breast cancer is used to infer the effect of radiation therapy on the development of a second cancer. While the outcome is observable for all subjects, the treatment (as well as some of the covariates) is missing for some of the subjects. Robins suggests assumptions on the missing data processes that imply point identification.

In practice, in applied work it is often assumed that the data are missing completely at random, and complete case (CC) analysis is conducted (see Little and Rubin (1987), Chapter 3, for a critical presentation of CC analysis). In this paper I examine the missing treatments problem from a "conservative" perspective, in the sense that I first determine the inferences that can be drawn in the absence of assumptions about the missing data mechanism and then illustrate the identifying power of widely credible assumptions posed on the distribution of the missing treatments and on the treatment selection rule.

In the setting considered here there are three problems preventing point identification of treatment effects: the usual latent outcome problem, the impossibility of identifying the distribution of received treatments, and the impossibility of matching the unobserved received treatments with the observed outcomes. I will propose a method to jointly address these problems. As will appear from the analysis of the following sections, significant progress can be made when the researcher has some prior information on the distribution of the missing treatments.

The paper is organized as follows: Section 2 introduces the relevant notation and the questions of interest. Section 3 introduces the empirical illustration, which focuses on the effect of drug abuse during work hours on unemployment, and shows that the assumption of ignorability of the observations with missing treatment data can be rejected for this problem. Section 4 derives worst case bounds on the treatment effects of interest. I show that in observational studies, when nothing is assumed about the distribution of the missing treatments, no information can be extracted from the observations for which the received treatment is unknown, and the additional degree of underidentification is proportional to the fraction of missing data. However, if some prior knowledge of the marginal distribution of the unobservable received treatments is available, some progress can be made. In Section 5 I show that under the maintained assumption of monotone treatment response (Manski (1997)), such prior knowledge is not necessary to extract information from the observations for which the received treatment is unknown. The bounds derived in this paper are sharp, in the sense that they exhaust all information available from the data and the maintained assumptions. In Sections 3-5, to keep the focus on identification, I treat identified quantities as known. Section 6 addresses statistical considerations by deriving two types of confidence intervals: a type that asymptotically covers the identification regions with a prespecified probability, following the approach of Chernozhukov, Hong, and Tamer (2004), and a type that asymptotically covers the true parameter of interest (rather than its identification region) with at least a prespecified

probability, following the approach of Imbens and Manski (2004). Section 7 presents the results of the empirical illustration. Section 8 concludes.

## 2 Setup of the Problem

Using standard notation (e.g., Neyman (1923)), let each member $j$ of a population of interest $J$ be characterized by some covariates $x_j \in X$, be exposed to a set of mutually exclusive and exhaustive treatments $T$, and have a specific response function $y_j(\cdot) : T \to Y$ mapping treatments $t \in T$ into outcomes $y_j(t) \in Y$. If $z_j \in T$ is the treatment that individual $j$ actually receives, then $y_j \equiv y_j(z_j)$ is the realized outcome, while $y_j(t)$ is a latent outcome for $t \neq z_j$. Denote by $d_j$ a binary variable which takes value 1 if the treatment received by individual $j$ is observed, 0 otherwise, and assume that the population is a measure space $(J, \Omega, P)$, with a probability measure $P$. In the analysis developed in this paper I assume perfect observability of realized outcomes as well as covariates; I also assume that all variables are correctly measured. Such assumptions are maintained in order to focus attention on the problem of missing treatments; Horowitz and Manski's (1995, 1998 and 2000) results can be easily incorporated in the analysis in case of missing or contaminated outcome data, or missing covariate data. Molinari's (2005) results can be applied in the presence of classification error in the outcome data, the covariate data, or the treatment data.

The researcher learns the distribution $P[y, x, d]$ of realized outcomes, covariates, and observability of realized treatments, and the distribution $P[z|x, y, d = 1]$ of realized treatments given covariates, realized outcome and observability of the realized treatment. The researcher's problem is to learn the distribution $P[y(\cdot)|x]$ of response functions, in order to infer the effect of a treatment.

I study the problem of missing treatments in the context of observational studies; however this problem can appear as well in randomized experiments, either because the assigned treatments are partially unobservable, or because there is uncertainty about the degree of compliance of the individuals with the treatments. If there is full compliance and partial unobservability of the assigned treatment, it is easy to show that the problem can be expressed as a case of missing covariates identical to the one studied by Horowitz and Manski (1998), so their results apply.[1] If there is perfect observability of the assigned treatment but uncertainty on the degree of compliance, the problem can be approached adapting the method of Balke and Pearl (1994)-(1997). They consider the problem of an experimental study where random assignment has taken place, but compliance is not perfect (i.e. the treatment received differs from that assigned). In case of observability of the received treatments, Balke and Pearl (1994)-(1997) derive sharp bounds on the

[1] If the treatments are randomly assigned, the treatment effect of interest can be evaluated by means of a regression of realized outcomes on treatments.

average treatment effect by solving a complex linear programming problem. The same method can be used in case of unobservability of the received treatment or uncertainty about the degree of compliance, as long as we can perfectly observe $d$ (if it is only known that a fraction of the agents doesn't comply with the assigned treatments, but it is not known who does not comply, the method below does not apply, as we are facing a contamination problem).

To simplify notation I omit the covariates in all that follows. I assume that the outcome variable $y$ takes values in a bounded set $Y$, where $K_0 \equiv \inf Y$ and $K_1 \equiv \sup Y$ are known finite numbers. For ease of exposition I assume that $T = \{0, 1\}$; in case of multiple treatments, all of the results in Sections 4 and part of the results in Section 5 (to be specified below) still hold. Since the focus of this paper is on the missing treatments problem, I assume $0 < \Pr(d = 0) < 1$. The questions to be addressed are:[2]

- what can be learned about $E[y(t)]$, $t \in T$, the average outcome under a mandatory policy;
- what can be learned about $E[y(1)] - E[y(0)]$, the average treatment effect ($ATE$).

# 3 Empirical Application

## 3.1 The Effect on Unemployment of Drug Use During Work Hours

Illicit drug and alcohol abuse have generally been associated with huge economic costs, and such association has been a motivation for drug-related and alcohol-related public policies in the US and worldwide. A large share of these costs is related to reduced labor productivity. Harwood, Fountain, and Livermore (1998) report that in 1992 alcohol and drug abuse cost society an estimated $176.4 billion as a result of (i) lost productivity, due to premature death and illness among abusers, (ii) crime-related costs of abusers, (iii) time abusers spent in residential treatment, and (iv) developmental disabilities among fetal alcohol syndrome survivors. Shortfalls in productivity and employment among individuals with alcohol or drug abuse disorders accounted for estimated losses of $80.9 billion in lost productivity ($66.7 billion resulting from alcohol problems and $14.2 billion from drug problems).[3] Harwood, Fountain, and Livermore (1998) report as well that in the subpopulation of enrollees in publicly funded treatment providers, chronic and severe alcohol and drug abusers appear often to have great difficulties obtaining and keeping stable employment. Those who do have a job may often be intoxicated or high and unable to work, and commonly move from one part-time job to another.

[2] Formal results on what can be learned about $E[y(t)] - E[y(z)]$, i.e. the status quo treatment effect ($STE$), are available from the author upon request.

[3] These estimates were constructed using the data from the National Longitudinal Alcohol Epidemiologic Survey, and the microsimulation techniques used in studies of the RAND Health Insurance Experiment.

It may seem logical to expect a negative relationship between the presence and severity of an alcohol and/or drug abuse problem and the employment status, labor supply, and wage rate of individuals. However, while there is evidence that "problem drinking" is associated with lower employment rates and greater unemployment,[4] the findings on the effects of drug use on labor market outcomes are still controversial. For example Kandel and Davies (1990), Kaestner (1991), Register and Williams (1992), and Gill and Michaels (1992) find that drug use is positively correlated with wages, even when one accounts for individual characteristics and endogeneity. At the same time, Kaestner (1994) and Buchmueller and Zuvekas (1998) find significant negative effect of drug use on employment or labor supply for males, but insignificant or positive effect for females (Kaestner (1994)). Kandel and Yamaguchi (1987) study the effect of drug use on job mobility, and show that drug use predicts job turnover and decreased tenure on the job; however their results suggest that these effects probably reflect the influence of preexisting differences among individuals who start using drugs instead of the effects of drugs themselves.

One of the problems which affect the empirical work in this area is the relatively high fraction of missing data (see for example Pepper (2001)); people are reluctant to answer questions relative to illicit activities or stigmatized activities, like drug abuse or "problem drinking". In practice, researchers often assume that the data are missing completely at random (MCAR), and conduct their analysis only on the subpopulation of respondents. I adopt a more conservative approach and study the effect for the population as a whole of drug use during work hours on the probability of being fired, discharged or laid off.

### 3.2 Data

I use data from the National Longitudinal Survey of Youth (NLSY). In its base year of 1979 the NLSY interviewed 12,686 persons who were between the age of 14 to 22 at that time. The survey has been updated each year since 1979 (and every two years since the early 1990's). The data contain detailed information on a respondent's labor market experience, and family and personal background. Approximately half of the total NLSY respondents were randomly sampled, the remaining being selected to overrepresent certain demographic groups (see BLS (1999)). In all that follows I restrict attention to the randomly sampled subpopulation; hence, problems connected with sampling design can be ignored. In 1984 and 1988 the respondents were asked questions about their lifetime and current use of several illicit drugs. In 1984, a (randomly sampled) group of 1,441 respondents who had been employed either in that year or in the past were asked whether (on their most recent job since the 1983 interview) they had been under the effect of illicit drugs during work hours.

[4] See for example Mullahy and Sindelar (1995).

While self reports may provide a great deal of information about an individual's behavior, the validity of such reports is sometimes questioned. Mensch and Kandel (1998) compare the declared illicit drugs use in the 1984 youth survey with the reports of other national surveys of drug use, and find underreporting of drug abuse (other than marijuana). But they as well suggest that such underreporting seems to be more common among light drug users than heavy users, and more common among blacks and Hispanics than among whites. As Gleason, Veum, and Pergamit (1991) argue, it may be that individuals who use drugs at work are more frequent users, and as Mensch and Kandel (1998) document less likely to underreport their drug use. Embracing this argument, in the analysis which follows I will assume that whenever an individual chooses to answer the question on drug use during work hours, she/he answers the truth.[5] However, as drug use at work may be a highly socially unacceptable activity, I will allow for the decision to not respond to the questions to be motivated by respondents' reluctance to report that they did engage in such activity.

I will focus attention on the 1345 (randomly sampled) respondents who were employed in 1983. Thus my empirical analysis concerns the subpopulation of persons who, in the notation introduced in Section 2 (and which will be left implicit in Sections 4-7), have the shared observable covariate $x = \{\text{employed in 1983}\}$. Additional covariates will be briefly considered in the next subsection.

Out of this group, 236 persons answered that they did use drugs during work hours, 994 answered that they did not, and 115 refused to answer the question (8.55%). I take the outcome of interest to be the number of weeks an individual was unemployed during the calendar year 1983-84. If the respondent answered the question on drug abuse ($d_j = 1$), we observe whether she has been under the effect of drugs during work hours ($z_j = 1 | d_j = 1$) or not ($z_j = 0 | d_j = 1$); if the respondent skipped the question, we register the missing data ($d_j = 0$). In the subpopulation of respondents, the average number of weeks of unemployment is relatively low regardless of drug use. Abstracting from sample variability, $E[y|\, z = 1, d = 1] = 3.267$, and $E[y|\, z = 0, d = 1] = 2.703$, while $\Pr[z = 1|\, d = 1] = 0.192$. If one were to assume random treatment selection (i.e. $z \perp \{y(0), y(1)\}$) and ignorability of the observations with missing treatment data (i.e. $d \perp \{y(0), y(1), z\}$), he would conclude that the average treatment effect is equal to 0.564 weeks. This would imply that drug use increases the average number of weeks of unemployment in a calendar year by little over half a week. Table 1 summarizes these descriptive statistics, along with their 95% confidence intervals.

### 3.3 Testing for MCAR and for Validity of the Complete Case Analysis

When facing missing data problems as the one described above, researchers often conduct "complete case" analyses, in which observations with any missing values are simply discarded (Little (1992)).

[5] This assumption can be relaxed, allowing for misreporting of drug use during work hours, using the direct misclassification approach of Molinari (2005).

Table 1: NLSY Descriptive Statistics, N=1345

| | Point Est. | 95% CI |
|---|---|---|
| Probability of missing treatments: $\Pr[d=0]$ | 0.086 | [0.071 , 0.100] |
| Probability of drug use given observable treatments: $\Pr[z=1\|d=1]$ | 0.192 | [0.170 , 0.214] |
| Average number of weeks of unemployment in subpopulation for whom treatments are observed and drugs are used: $E[y\|z=1,d=1]$ | 3.267 | [2.322 , 4.212] |
| Average number of weeks of unemployment in subpopulation for whom treatments are observed and drugs are not used: $E[y\|z=0,d=1]$ | 2.703 | [2.303 , 3.103] |
| Average number of weeks of unemployment in subpopulation for whom treatments are observed: $E[y\|d=1]$ | 2.811 | [2.441 , 3.182] |
| Average number of weeks of unemployment in subpopulation for whom treatments are not observed: $E[y\|d=0]$ | 1.930 | [1.502 , 2.359] |
| Difference in the average number of weeks of unemployment between subpopulation for whom treatments are observed and for whom treatments are unobserved: $E[y\|d=1]-E[y\|d=0]$ | 0.881 | [0.314 , 1.448] |
| $ATE$ assuming exogenous treatment selection and treatments missing completely at random: $E[y\|z=1,d=1]-E[y\|z=0,d=1]$ | 0.564 | [−0.462 , 1.590] |

In the context of treatment effects, when the treatment data are partially unobservable, this can lead to valid inference if: (i) the treatment data are MCAR, (ii) $d \perp \{(y(0), y(1))|\, x\}$, (iii) $d \perp \{(y(0), y(1), z)|\, x\}$. Note that assuming $P[y(t)|\, z, d, x] = P[y(t)|\, z, x]$ alone does not imply validity of the CC analysis. Validity would be assured if one assumed also that $\Pr(z = t|\, d, x) = \Pr(z = t|\, x)$. The assumption $d \perp \{(y(0), y(1))|\, x\}$, is not testable, and in general it does not seem appealing. For example, if $\Pr(z = t|\, d, x) \neq \Pr(z = t|\, x)$ and $P[y(t)|\, z, d, x] = P[y(t)|\, z, x]$, it follows that $P[y(t)|\, x, d = 1] \neq P[y(t)|\, x, d = 0]$. Similarly, if $\Pr(z = t|\, d, x) = \Pr(z = t|\, x)$ and $P[y(t)|\, z, d, x] \neq P[y(t)|\, z, x]$, it follows again that $P[y(t)|\, x, d = 1] \neq P[y(t)|\, x, d = 0]$.

On the other hand, a necessary condition for the assumption $d \perp \{(y(0), y(1), z)|\, x\}$, as well as for the assumption MCAR, to hold can be tested. In particular, the following result is easy to verify:

**Lemma 1** *Suppose that* $d \perp \{(y(0), y(1), z)|\, x\}$. *Then*

$$P[y|\, d = 1, x] = P[y|\, d = 0, x] \tag{1}$$

Clearly, if the observability of the treatments is independent from the response functions and from the received treatments conditional on the observed covariates, the distribution of realized

outcomes, conditional on the observed covariates, for the subpopulation for whom the received treatments are observable should be the same as that for the subpopulation with missing treatment data. In practice, before conducting CC analyses, one can test whether the equality in (1) holds. Note that if the assumptions $P[y(t)|z,d,x] = P[y(t)|z,x]$ and $P(z|d,x) = P(z|x)$ are jointly maintained, condition (1) should again hold, and therefore the same test as described in Lemma 1 can be performed. However in both cases, condition (1) is only necessary but not sufficient for the CC analysis to be valid, or for the MCAR assumption to hold.

Given the NLSY sample, for $x = \{\text{employed in 1983}\}$ a Wilkoxon rank-sum test rejects the equality in (1) at the 5% significance level. Conditioning on additional covariates, the equality in (1) can again be rejected. For example, if we look at the group of respondents employed in 1983, younger than 24, with at least a high school degree, we can again reject the null at 5% significance level.

# 4 Worst Case Bounds

The following Subsections 4.1-4.2 analyze what can be learned about the treatment effects of interest when nothing is assumed about the distribution of the missing treatments. The availability of prior information on this distribution will turn out to be crucial: if nothing is known about $\Pr[z=1|d=0]$, no information can be extracted from the observations for which the treatments are missing. However, if a bound (or point identification) on $\Pr[z=1|d=0]$ is available, it is possible to extract information from the observations with missing treatment data; Subsection 4.3 discusses how.

## 4.1 Worst Case Bounds on a Mandatory Policy

Suppose that we are interested in $E[y(1)]$; by the Law of Iterated Expectations

$$E[y(1)] = E[y(1)|d=1]\Pr(d=1) + E[y(1)|d=0]\Pr(d=0) \tag{2}$$

From the data we can learn $\Pr(d=1)$ and $\Pr(d=0)$; regarding the other two terms on the right hand side of (2), let us first focus our attention on $E[y(1)|d=1]$:

$$E[y(1)|d=1] = E[y|d=1,z=1]\Pr[z=1|d=1] + E[y(1)|d=1,z=0]\Pr[z=0|d=1] \tag{3}$$

The equality in (3) expresses the usual problem of learning the distribution of outcomes under a mandatory policy (in this case assigning treatment 1): the only unobserved quantity is the counterfactual probability of success under the treatment for people who actually did not receive it: $E[y(1)|d=1,z=0]$.

Let us now consider $E[y(1)|\, d=0]$:

$$E[y(1)|\, d=0] = E[y|\, d=0, z=1]\Pr[z=1|\, d=0] + E[y(1)|\, d=0, z=0]\Pr[z=0|\, d=0] \quad (4)$$

The problem arising in the case of missing treatments is that all quantities in (4) are unknown: not only do we have the usual problem of latent outcomes, but we also do not know the distribution of treatments when they are unobservable; moreover the data do not reveal how to match the realized outcome with the received but unobservable treatment. The only thing we can learn from the data is the distribution of realized outcomes under unobserved treatments, i.e. we can learn

$$Q(t) \equiv \Pr[y \leq t|\, d=0] = \Pr[y \leq t|\, d=0, z=1]\, p + \Pr[y \leq t|\, d=0, z=0]\,(1-p)$$

where $p \equiv \Pr[z=1|\, d=0]$. If $p$ is known, it is possible to use the result of Corollary 4.1 in Horowitz and Manski (1995) to find a sharp bound on $E[y(1)|\, d=0]$ using the information provided by $Q$ (similarly for $E[y(0)|\, d=0]$). However, if we do not know anything about $p$, knowledge of $Q$ does not help to bound the outcome distribution under received treatment unobservability. Indeed, it may be that all treatments we do not observe are of type 0, in which case knowledge of $Q$ does not provide any information about $E[y(1)]$.

Proposition 1 states what can be learned about a mandatory policy, both when $p$ is known and when $p$ is unknown. Before stating the result, I need to introduce some additional notation: for any $\pi \in [0,1]$, let $r(\pi)$ denote the $\pi-$quantile of $Q$:

$$r(\pi) \equiv \inf_t \{Q(t) \geq \pi\}.$$

Define probability distributions $L_\pi$ and $U_\pi$ on $\Re$ as follows

$$\begin{aligned} L_\pi[-\infty, t] &\equiv \frac{Q(t)}{\pi} && \text{for } t < r(\pi), \\ &\equiv 1 && \text{for } t \geq r(\pi). \end{aligned}$$

$$\begin{aligned} U_\pi[-\infty, t] &\equiv 0 && \text{for } t < r(1-\pi), \\ &\equiv \frac{Q(t) - (1-\pi)}{\pi} && \text{for } t \geq r(1-\pi). \end{aligned}$$

**Proposition 1** *Given the value of $p \in [0,1]$ and no other information, the sharp bounds on $E[y(1)|\, d=0]$ and on $E[y(1)]$ are given by:*

$$E[y(1)|\, d=0] \in \left[p\textstyle\int y dL_p + (1-p)K_0,\, p\int y dU_p + (1-p)K_1\right] \quad (5)$$

$$\begin{aligned} &\left(E[y|\, d=1, z=1]\Pr[z=1, d=1] + P[z=0, d=1]K_0\right) + \left(p\textstyle\int y dL_p + (1-p)K_0\right)\Pr(d=0) \\ &\qquad \leq E[y(1)] \leq \\ &\left(E[y|\, d=1, z=1]\Pr[z=1, d=1] + P[z=0, d=1]K_1\right) + \left(p\textstyle\int y dU_p + (1-p)K_1\right)\Pr(d=0) \end{aligned} \quad (6)$$

*In absence of knowledge of p, the sharp bounds on* $E[y(1)|d=0]$ *and on* $E[y(1)]$ *are given by:*

$$K_0 \leq E[y(1)|d=0] \leq K_1 \tag{7}$$

$$E[y|d=1,z=1]\Pr[z=1,d=1] + (P[z=0,d=1] + \Pr(d=0))K_0 \leq$$
$$E[y(1)] \leq E[y|d=1,z=1]\Pr[z=1,d=1] + (P[z=0,d=1] + \Pr(d=0))K_1$$

**Proof.** See Appendix. ■

While the width of the bound on $E[y(1)]$ with no missing data is equal to $\Pr[z=0](K_1-K_0) = \Pr[z=0|d=1](K_1-K_0)$, with missing treatments it increases to $(\Pr[z=0|d=1]\Pr(d=1) + \Pr(d=0))(K_1-K_0)$. This expression can be rewritten as

$$(\Pr[z=0|d=1] + \Pr[z=1|d=1]\Pr(d=0))(K_1-K_0).$$

It is then easy to see that the increase in the width is proportional to $\Pr[z=1|d=1]\Pr(d=0)$, and hence, for given fraction of individuals observed to receive treatment 1, proportional to the fraction of missing data.

### 4.2 Worst Case Bounds on the Average Treatment Effect

Consider now the case in which the researcher is interested in learning about the average treatment effect ($ATE$)

$$E[y(1)] - E[y(0)] \tag{8}$$

As we did for $E[y(1)]$, we can decompose the quantity in (8) using the Law of Iterated Expectations:

$$\begin{aligned} E[y(1)] - E[y(0)] &= \{E[y(1)|d=1] - E[y(0)|d=1]\}\Pr(d=1) \\ &\quad + \{E[y(1)|d=0] - E[y(0)|d=0]\}\Pr(d=0) \end{aligned}$$

Let us first consider the $ATE$ under observability of the received treatment. A sharp bound on this quantity can be found using results well known in the literature (e.g., Manski (1995)):

$$LB_{ATE}^{d=1} \leq E[y(1)|d=1] - E[y(0)|d=1] \leq UB_{ATE}^{d=1} \tag{9}$$

where

$$\begin{aligned} LB_{ATE}^{d=1} = \; & E[y|d=1,z=1]\Pr[z=1|d=1] + K_0\Pr[z=0|d=1] \\ & -E[y|d=1,z=0]\Pr[z=0|d=1] - K_1\Pr[z=1|d=1] \end{aligned} \tag{10}$$

$$\begin{aligned} UB_{ATE}^{d=1} = \; & E[y|d=1,z=1]\Pr[z=1|d=1] + K_1\Pr[z=0|d=1] \\ & -E[y|d=1,z=0]\Pr[z=0|d=1] - K_0\Pr[z=1|d=1] \end{aligned} \tag{11}$$

Note that the width of this bound is exactly $(K_1 - K_0)$, as for the usual $ATE$.

Consider now the $ATE$ under unobservability of the received treatment. In what follows let

$$\begin{array}{ll} \Pr[z=1|\, d=0] \equiv p & P[y|\, d=0] \equiv Q \\ E[y|\, d=0, z=1] \equiv E_{11} & E[y(1)|\, d=0, z=0] \equiv E_{10} \\ E[y|\, d=0, z=0] \equiv E_{00} & E[y(0)|\, d=0, z=1] \equiv E_{01} \end{array}$$

Then we can restate the problem as to find a bound on

$$E[y(1)|\, d=0] - E[y(0)|\, d=0] = E_{11}p + E_{10}(1-p) - E_{00}(1-p) - E_{01}p \tag{12}$$

Note that all quantities on the right hand side of (12) are unknown; however, for a given value of $p$ we can extract information from the knowledge of $Q$. As $p$ ranges from 0 to 1 we move from knowing exactly the value of $E_{11}$, to having decreasing information on it and increasing information on $E_{00}$, to knowing exactly the value of $E_{00}$. Even if $E_{01}$ and $E_{10}$ are unknown, using this fact we can find a sharp $p$-dependent bound on the average treatment effect under treatment unobservability.

**Proposition 2** *Given the value of $p \in [0,1]$ and no other information, the sharp bound on $E[y(1)|\, d=0] - E[y(0)|\, d=0]$ is given by:*

$$\begin{array}{rl} LB_{ATE}^{d=0} = & p\int y dL_p + (1-p)K_0 - (1-p)\int y dU_{1-p} - pK_1 \\ UB_{ATE}^{d=0} = & p\int y dU_p + (1-p)K_1 - (1-p)\int y dL_{1-p} - pK_0 \end{array} \tag{13}$$

*For any value of $p$ this bound is informative. The sharp bound on $E[y(1)] - E[y(0)]$ is given by:*

$$\begin{array}{rcl} LB_{ATE}^{d=1}\cdot\Pr(d=1) + LB_{ATE}^{d=0}\cdot\Pr(d=0) & \leq & E[y(1)] - E[y(0)] \\ & \leq & UB_{ATE}^{d=1}\cdot\Pr(d=1) + UB_{ATE}^{d=0}\cdot\Pr(d=0) \end{array} \tag{14}$$

*where $LB_{ATE}^{d=1}$ and $UB_{ATE}^{d=1}$ were defined in (10)-(11).*
*In absence of knowledge of $p$, the sharp bounds on $E[y(1)|\, d=0] - E[y(0)|\, d=0]$ and on $E[y(1)] - E[y(0)]$ are given by:*

$$-(K_1 - K_0) \leq E[y(1)|\, d=0] - E[y(0)|\, d=0] \leq (K_1 - K_0) \tag{15}$$

$$\begin{array}{rcl} LB_{ATE}^{d=1}\cdot\Pr(d=1) - (K_1-K_0)\Pr(d=0) & \leq & E[y(1)] - E[y(0)] \\ & \leq & UB_{ATE}^{d=1}\cdot\Pr(d=1) + (K_1-K_0)\Pr(d=0). \end{array} \tag{16}$$

*The width of the band in (16) is equal to $(K_1 - K_0)(1 + \Pr(d=0))$.*

**Proof.** See Appendix. ■

### 4.3 Identifying Power of Assumptions on $\Pr[z=1|d=0]$

The bounds derived in Propositions 1 and 2 are functions of $p \equiv \Pr[z=1|d=0]$. We showed that as long as $p$ is unknown no information can be extracted from the observations with missing treatment data. However, if some prior information on the distribution of the missing treatments is available, some progress can be made.[6]

Suppose for example that $p$ is identified, either because the results of a validation study are available, or because one is willing to assume that the treatments are missing at random,[7] i.e. $d \perp z$, which implies $p = \Pr[z=1|d=1]$. Then the bounds reported in Propositions 1 and 2 can be evaluated at this value of $p$.

Alternatively, one may learn that $p \in [p_1, p_2]$. For example, if the survey from which the data are drawn concerns activities to which a stigma is associated (e.g.: illicit drug use), the decision not to respond to the questions can be motivated by respondents' reluctance to report that they engaged in such activities. In this case, it may be credible to assume that the probability of having engaged in such activities for people who didn't answer the survey is not smaller than for people who did answer the question. Pepper (2001) introduces an assumption of this type. Then the bound on the $ATE$ can be calculated by taking (respectively) the infimum value and the supremum value that the lower bound and the upper bound reported in Proposition 2 achieve for $p$ ranging in $[p_1, p_2]$. Regarding the mandatory policy, the bound on $E[y(1)]$ can be calculated by evaluating the bound in Proposition 1 at $p = p_1$ (while the information provided by $p_2$ can be used to bound more tightly $E[y(0)]$).[8]

## 5 Missing Treatments with Monotonicity Assumptions

Manski (1997) investigates what may be learned about treatment response when it is assumed that response functions are monotone, semi-monotone, or concave monotone, and no assumptions are imposed on the treatment selection process. He shows that assuming monotone or concave monotone response qualitatively improves the identification problem relative to the worst case situation in which no prior information is available. Manski and Pepper (2000) study the identifying power of monotone instrumental variable assumptions, and in particular the special case of monotone treatment selection ($MTS$). They show that the joint assumption of monotone treatment selection

[6] A more detailed derivation of the identifying power of assumptions on $P[z=1|d=0]$ is available from the author upon request.

[7] Note that in this case we are not assuming $d \perp (y(0), y(1), z)$.

[8] Another example along these lines is given by recent work of Kreider and Hill (2005), who apply some of the results in this paper to study the problem of learning utilization rates of health services under a hypothetical policy of universal health insurance, when insurance status is subject to classification error and is unverified for some respondents (the respondents with missing treatments).

and response can be tested, and that if such joint assumption is maintained, informative bounds are obtainable even if $Y$ is unbounded.

In this section I study the identifying power of assuming monotone treatment response ($MTR$), and of jointly assuming monotone treatment response and selection ($MTR-MTS$), when some of the treatments are missing. I show that under the maintained assumption of monotone treatment response (alone or jointly with $MTS$), one can extract information from the observations for which the treatment data are missing even without any prior knowledge of $p \equiv P[z=1|\, d=0]$. Clearly, prior information on $p$ will further shrink the bounds.

Let us assume that the response function is weakly increasing; in the case studied here, since $T=\{0,1\}$, this implies that for each $j \in J$:

$$y_j(0) \leq y_j(1) \tag{17}$$

To interpret this assumption, consider the example of the effect of drug abuse during work hours on the number of weeks an individual is unemployed in a calendar year. The monotone treatment response assumption implies that each person's unemployment outcome is weakly increasing in conjectured use of drugs during work hours.

If (17) holds, we can use the results of Manski (1997) to tighten the bounds on the quantities of interest given observability of the received treatment; in particular we get the following sharp bounds:

$$K_0 \Pr[z=1|\, d=1] + E[y|\, z=0, d=1] \Pr[z=0|\, d=1] \leq E[y(0)|\, d=1] \leq E[y|\, d=1] \tag{18}$$

$$E[y|\, d=1] \leq E[y(1)|\, d=1] \leq E[y|\, z=1, d=1] \Pr[z=1|\, d=1] + K_1 \Pr[z=0|\, d=1] \tag{19}$$

Under the $MTR$ assumption we can as well extract information from knowledge of the distribution of realized outcomes given unobservability of the treatment, and get tighter bounds on $E[y(0)|\, d=0]$ and $E[y(1)|\, d=0]$. Then, through the Law of Iterated Expectations and given (18)-(19), we get narrower bounds on $E[y(0)]$ and $E[y(1)]$.

**Proposition 3** *Suppose that the treatment response function is weakly increasing. Then:*

$$\begin{array}{l} K_0 \leq E[y(0)|\, d=0] \leq E[y|\, d=0] \\ E[y|\, d=0] \leq E[y(1)|\, d=0] \leq K_1 \end{array} \tag{20}$$

*and*

$$\begin{array}{rcl} K_0 \Pr[z=1, d=1] + E[y|\, z=0, d=1] \Pr[z=0, d=1] + K_0 \Pr(d=0) & \leq & \\ E[y(0)] & \leq & E(y) \end{array} \tag{21}$$

$$\begin{aligned} E(y) &\leq E[y(1)] \\ &\leq E[y|\, z=1, d=1]\Pr[z=1, d=1] + K_1 \Pr[z=0, d=1] + K_1 \Pr(d=0) \end{aligned} \tag{22}$$

$$\begin{gathered} 0 \leq E[y(1)] - E[y(0)] \leq \\ \left\{ \begin{array}{c} (E[y|\, z=1, d=1] - K_0)\Pr[z=1, d=1] \\ +(K_1 - K_0)\Pr(d=0) + (K_1 - E[y|\, z=0, d=1])\Pr[z=0, d=1] \end{array} \right\} \end{gathered} \tag{23}$$

*In absence of additional information, these bounds are sharp.*

**Proof.** See Appendix. ■

The result in (20) is quite intuitive: if we assume $y_j(0) \leq y_j(1)$, it follows that assigning treatment 0 as a mandatory policy can't imply a larger outcome than the realized one, while assigning treatment 1 as a mandatory policy can't imply a smaller outcome than the realized one.

In case of multiple treatments, i.e. $T = \{t_1, t_2, \ldots, t_N\}$ and $t_1 \leq t_2 \leq \ldots \leq t_N$, one can use the results in (20) to get tighter bounds on $E[y(t_1)|\, d=0]$ and on $E[y(t_N)|\, d=0]$. However no information can be extracted from the observations with incomplete data in order to tighten the bounds on $E[y(t)|\, d=0]$ for $\forall\ t \in T, t \neq t_1, t_N$, unless prior information is available on $\Pr[z=1|\, d=0]$. Regarding $E[y(t)|\, d=0] - E[y(s)|\, d=0]$, under the $MTR$ Assumption this quantity ranges in $[0, K_1]$ for $\forall\ t, s \in T$, $t > s$.

Let us now assume that $E[y(t)|\, z=u]$ is weakly increasing in $u$; let's assume further that this restriction holds for the subpopulation with complete data. In the case studied here, since $T = \{0, 1\}$, this implies that for each $t \in T$:

$$E[y(t)|\, z=1, d=1] \geq E[y(t)|\, z=0, d=1] \tag{24}$$

To interpret this assumption, consider the example of the effect of drug abuse during work hours on the average number of weeks of unemployment. The monotone treatment selection assumption implies that, for each $t \in T$, persons who do not select into using drugs during work hours experience a weakly lower average number of weeks of unemployment than do people who do select into using drugs during work hours. Note that I require such assumption to hold for the subpopulation for which treatments are observable since to make use of it one needs to observe the received treatments. This implies that the $MTS$ Assumption does not have identifying power on $E[y(t)|\, d=0]$, unless prior information is available on $\Pr[z=1|\, d=0]$.

Given (24), we can use the results of Manski and Pepper (2000) to tighten the bounds on the quantities of interest given observability of the received treatment; in particular we get the following sharp bounds:

$$E[y|\, z=0, d=1] \leq E[y(0)|\, d=1] \leq E[y|\, z=0, d=1]\Pr[z=0|\, d=1] + K_1 \Pr[z=1|\, d=1]$$

$$\begin{aligned} K_0 \Pr[z=0|d=1] + E[y|z=1,d=1]\Pr[z=1|d=1] &\leq E[y(1)|d=1] \\ &\leq E[y|z=1,d=1] \end{aligned}$$

If one is willing to impose jointly the $MTR$ and $MTS$ assumptions, the width of the bounds shrinks significantly. In particular, for the subpopulation with no missing data, under the joint $MTR-MTS$ Assumption

$$E[y|z=0,d=1] \leq E[y(0)|d=1] \leq E[y|d=1] \tag{25}$$

$$E[y|d=1] \leq E[y(1)|d=1] \leq E[y|z=1,d=1] \tag{26}$$

$$0 \leq E[y(1)|d=1] - E[y(0)|d=1] \leq E[y|z=1,d=1] - E[y|z=0,d=1] \tag{27}$$

The following Proposition, which can be easily verified given the results in (20), the results in (25)-(27), and the Law of Iterated Expectations, shows what are the bounds on the treatment effects of interest under the joint $MTR-MTS$ Assumption.

**Proposition 4** *Suppose that (17) and (24) jointly hold. Then:*

$$E[y|z=0,d=1]\Pr(d=1) + K_0 \Pr(d=0) \leq E[y(0)] \leq E(y)$$

$$E(y) \leq E[y(1)] \leq E[y|z=1,d=1]\Pr(d=1) + K_1 \Pr(d=0)$$

$$\begin{aligned} 0 &\leq E[y(1)] - E[y(0)] \\ &\leq \{E[y|z=1,d=1] - E[y|z=0,d=1]\}\Pr(d=1) + (K_1 - K_0)\Pr(d=0) \end{aligned}$$

*In absence of additional information, these bounds are sharp.*

The joint $MTR-MTS$ Assumption is a testable hypothesis, which should be rejected if $E[y|z=u]$ is not weakly increasing in $u$. When some of the treatments are missing, it is straightforward to show that such assumption can still be tested on the subpopulation for which the treatments are observable. However the hypothesis cannot be tested on the subpopulation with missing data, and hence on the population as a whole.

# 6 Confidence Sets for the Parameters of Interest and for the Identification Regions

Confidence intervals which asymptotically cover the identification regions derived in Sections 4-5 with a prespecified probability $(1-\alpha)$ can be obtained by using the results of Chernozhukov, Hong,

and Tamer (2004). A conceptually different type of confidence regions that asymptotically cover the true parameter of interest (rather than its identification region) with probability at least $(1-\alpha)$ can be obtained by using the results of Imbens and Manski (2004). In both cases, denoting by $\hat{\vartheta}_L$ and $\hat{\vartheta}_U$ the estimated lower and upper bound for a certain parameter of interest, the confidence interval is of the form $\left[\hat{\vartheta}_L - c_{L\alpha}, \hat{\vartheta}_U + c_{U\alpha}\right]$, where $c_{L\alpha}$ and $c_{U\alpha}$ depend on the critical value of a certain distribution, which differs in the two approaches, and on the sample size. In this paper I utilize both methodologies, and compare the confidence intervals obtained for the empirical application in Section 7.

Chernozhukov, Hong and Tamer (2004) discuss how to construct confidence sets for identification regions of parameters obtained as the solution of the minimization of a criterion function. They also discuss how to construct confidence regions when the object of interest is a parameter of the form $E[y(1)]$ or $E[y(1)-y(0)]$, the identification region for such parameter is the entire interval between two functionals of the distribution of the observed data ($\hat{\vartheta}_L$ and $\hat{\vartheta}_U$), and the estimates of the lower and upper bounds converge at the parametric rate. In order to be able to use their result, I need to derive the joint asymptotic distribution of the sample analogs of each of the lower and upper bounds obtained in Sections 4-5. Using standard arguments on the asymptotic properties of sample means and L-statistics (Shorack and Wellner (1986), Chapter 19), I provide these results in Appendix B.

The confidence intervals proposed by Imbens and Manski (2004) asymptotically cover the true parameter of interest with at least a prespecified probability. They are designed for intervally identified parameters, and require the derivation of the joint asymptotic distribution of the sample analogs of the lower and upper bounds of the parameter of interest. However, in order to ensure uniformity in the coverage properties of their confidence intervals (and to avoid anomalies in the width of the confidence intervals that arise as one approaches point identification), Imbens and Manski (2004) require that a uniform central limit theorem hold for the joint asymptotic distribution of the sample analogs of each of the lower and upper bounds of interest. Such requirement is satisfied for the lower and upper bounds in Section 5, and for some of those in Section 4. However, some of the endpoints of the bounds obtained in Sections 4 involve randomly trimmed means, and, to the best of my knowledge, uniform CLT for such estimators are not available, unless the fraction of trimmed observations converges to zero as the sample size goes to infinity (Shorack (1997)). Alternatively, Berry-Esseen results are available for trimmed means, provided the trimming proportion is non-random (see, e.g., Serfling (1984)).[9] Since the lower and upper bounds in Section 4 do not satisfy either of these requirements, uniformity of the convergence of the confidence intervals to their nominal level is not guaranteed, and is the subject of ongoing research.

[9] Chavez-Martin-Del-Campo (2005) uses these results to derive Imbens-Manski confidence intervals for the case of identification and estimation of poverty measures with contaminated data.

# 7 Results for the Empirical Application

Tables 2-3 report respectively the bounds on the average outcome under a mandatory policy of no illicit drug use, and the bounds on the $ATE$, along with the Imbens and Manski (IM) and the Chernozhukov, Hong and Tamer (CHT) 95% confidence intervals, for the five different sets of assumptions considered in the previous sections. As predicted by the theory, the CHT confidence intervals are always larger than the IM confidence intervals, as the former cover the entire identification region with a prespecified probability, while the latter cover the true parameter of interest. In practice, in this empirical application the difference in the width of the confidence intervals is relatively small. The biggest difference for the case of the average outcome under a mandatory policy of no illicit drug use is 13% of the width of the corresponding IM confidence intervals, and occurs under the $MTR-MTS$ assumption. The biggest difference for the case of the ATE is 3% of the width of the corresponding IM confidence intervals, and occurs again under the $MTR-MTS$ assumption.

Since the outcome of interest in this application is the number of weeks an individual is unemployed in a calendar year, I set $K_0 = 0$ and $K_1 = 52$. The first set of bounds in each Table ("No Assumptions") constitute the baseline of our analysis. With no assumptions on the distribution of the missing treatments and on the treatment selection rule the bounds are necessarily wide; this is mainly due to the non observability of latent outcomes, which for example implies that the bound on the $ATE$ is at least of width $(K_1 - K_0) = 52$. The second set of bounds in each Table ("Complete Cases") assumes that the observations with missing treatments data can be ignored. To interpret the difference in the results between the first and second row in Table 2 and in Table 3, notice that under the maintained assumption of ignorability of the observations with missing treatments, only the subsample of observations with complete data is used ($N = 1230$). On the other hand, when nothing is assumed about the distribution of the missing treatments, all observations are used to draw inference ($N = 1345$). However, as long as $p \equiv P[z = 1|\, d = 0]$ is unknown, no information can be extracted from the subsample with missing treatment data.

As 80.8% of the respondents report that they have not been under the effect of drugs during work hours, the bounds on the average number of weeks of unemployment under a mandatory policy of no illicit drug use are relatively narrow. Notice that the bound is much wider if nothing is assumed about the distribution of the missing treatments, in which case the width is 13.57 weeks, than if we assume ignorability of the observations with missing treatments data, in which case the width is 9.98 weeks (that is, the width increases by 36% once we account for the missing treatments). Regarding the bound on the average treatment effect ($ATE$), while ignoring the observations with missing treatments data the width of the bound is equal to 52 weeks, when nothing is assumed about the distribution of missing treatments the width enlarges to $(1 + \Pr(d = 0)) K_1 = 56.41$.

Table 2: Bounds and Confidence Intervals for the Average Number of Weeks of Unemployment under a Mandatory Policy of no Illicit Drug Use: E[y(0)]

| | Bounds | 95% IM CI | 95% CHT CI |
|---|---|---|---|
| No Assumptions | [1.998 , 15.568] | [1.744 , 16.572] | [0.995 , 16.571] |
| Complete Cases | [2.185 , 12.162] | [1.908 , 13.116] | [1.225 , 13.121] |
| $TMR$ | [1.998 , 14.715] | [1.744 , 15.592] | [1.123 , 15.590] |
| $MTR$ | [1.998 , 2.736] | [1.744 , 3.029] | [1.670 , 3.063] |
| $MTR - MTS$ | [2.472 , 2.736] | [2.161 , 3.031] | [2.112 , 3.096] |

Table 3: Bounds and Confidence Intervals for the ATE: E[y(1)] - E[y(0)]

| | Bounds | 95% IM CI | 95% CHT CI |
|---|---|---|---|
| No Assumptions | [−14.995 , 41.451] | [−15.967 , 42.295] | [−16.081 , 42.538] |
| Complete Cases | [−11.535 , 40.465] | [−12.443 , 41.374] | [−12.611 , 41.542] |
| $TMR$ | [−11.567 , 40.762] | [−12.478 , 41.669] | [−12.648 , 41.843] |
| $MTR$ | [ 0 , 41.451] | [ 0 , 42.295] | [ 0 , 42.454] |
| $MTR - MTS$ | [ 0 , 4.962] | [ 0 , 5.982] | [ 0 , 6.173] |

In order to illustrate the identifying power of assumptions on $P[z = 1|\, d = 0]$, the third row of Table 2 and the third row of Table 3 report the bounds and the 95% confidence intervals on the treatment effects of interest under the maintained assumption of treatments missing at random ($TMR$; note that we are not assuming $d \perp (y(0), y(1), z)$), which can be specified as follows:

$$TMR: \; P[z = 1|\, d = 0] = P[z = 1|\, d = 1]$$

This assumption states that respondents who did not answer the question relative to drug abuse are as likely as respondents who did answer the question to have been under the effect of drugs during work hours. This assumption can be credible if for example some of the people interviewed in the NLSY are "impatient". Suppose that a fraction of the respondents get easily bored answering the questionnaire, and hence don't answer some of the questions, including the one on drug use. Suppose further that drug abuse is independent from impatience; then we can assume that the treatments are missing at random. However, unemployment outcomes may be correlated with impatience, hence it may not be reasonable to assume that the outcomes' distribution is identical for patient and impatient people. Given the NLSY sample, this assumption implies that $P[z = 1|\, d = 0] = 0.192$.

Once $P[z = 1|\, d = 0]$ becomes known, it is possible to use the results in Propositions 1 and 2

to get narrower bounds on the average number of weeks of unemployment under mandatory policy of no illicit drug use, and on the $ATE$. Note that under the $TMR$ Assumption all observations are used to draw inference ($N = 1345$), and information can be extracted also from the subsample with missing treatment data, as illustrated in Section 4.

The last two rows of Tables 2–3 report the bounds and the confidence intervals on the treatment effects of interest, under the maintained assumption of monotone treatment response ($MTR$), and under the maintained assumption of joint monotone treatment response and selection ($MTR - MTS$). As already introduced in the previous section, such assumptions are specified as follows:

$$\begin{array}{ll} MTR: & y_j(1) \geq y_j(0) \\ MTS: & E[y(t)|\, z=1, d=1] \geq E[y(t)|\, z=0, d=1] \ , \ t \in \{0,1\} \end{array} \tag{28}$$

In this application, the monotone treatment response assumption states that each person's unemployment outcome is weakly increasing in conjectured use of drugs during work hours. The monotone treatment selection assumption states that, for all $t \in \{0,1\}$, persons who do not select into using drugs during work hours experience a weakly lower average number of weeks of unemployment than do people who do select into using drugs during work hours. In other words, the $MTS$ is a "sorting" assumption. In this application, it states that if we divide the population in two groups according to the received treatment, then the average outcome for the group who did not use drugs during work hours is lower than for the group who did use it, for each $t \in T$.

Manski and Pepper (2000) show that the joint $MTR{-}MTS$ Assumption is a testable hypothesis, which should be rejected if $E[y|\, z=u]$ is not weakly increasing in $u$. Table 1 reports that in the NLSY sample $E[y|\, z=1, d=1] = 3.267$ and $E[y|\, z=0, d=1] = 2.703$. The 95% confidence interval shows that we cannot reject the assumption that $E[y|\, z=1, d=1] - E[y|\, z=0, d=1] \geq 0$, and that the band contains everywhere monotone non-decreasing functions. Hence I proceed on the basis that $MTR - MTS$ assumption is consistent with the empirical evidence.

While the identifying power of the $MTR$ Assumption can be appreciated almost exclusively for the average number of weeks of unemployment under a mandatory policy of no illicit drug use, the joint $MTR - MTS$ Assumption appears to have substantial identifying power. In particular, if we look at $E[y(1)] - E[y(0)]$, under the maintained assumption of $MTR - MTS$ we conclude that using drugs during work hours can increase the average number of weeks of unemployment by up to 4.96 weeks. If we look at the 95% confidence intervals around the true parameter, we conclude that this increase can be of up to 5.98 weeks. If we look at the 95% confidence intervals around the bound, we conclude that this increase can be of up to 6.17 weeks.

# 8 Conclusions

This paper has introduced the problem of missing treatments, i.e. what we can learn from the data about a treatment effect of interest when we do not know which treatment some individuals have received. While for the observations for which we know the received treatment the problem is identical to the standard one, namely the non observability of the latent outcomes, when we do not observe the received treatments we face a new issue. Not only we have the usual latent outcome problem, but we do not know the distribution of the treatments, and we are also unable to match the unobserved received treatments with the observed realized outcomes.

I considered the missing treatments problem in the context of observational studies and of survey nonresponse relative to the variable to be used as a treatment, and I showed that the assumption of ignorability of the observations with missing treatment data can be tested. Sharp worst case bounds were derived for the case in which no assumptions are imposed; I showed that no information can be extracted from the observations for which the received treatment is unknown, and that the additional degree of under-identification is proportional to the fraction of missing data. I illustrated how to use prior information on the distribution of the missing treatments (as for example prior knowledge of $P[z = 1|\, d = 0]$ or bounds on this quantity) to shrink the width of the bounds. Finally, I showed that under the maintained assumption of monotone treatment response information can be extracted from the observations with missing treatment data even without any prior knowledge of $P[z = 1|\, d = 0]$.

The theoretical results were illustrated by means of an empirical application studying the effect of drug abuse during work hours on unemployment. Given the NLSY sample, I showed that under the maintained assumption of joint monotone treatment response and selection, using drugs during work hours can increase the average number of weeks of unemployment by up to 4.96 weeks.

# A Proofs of Propositions

In all that follows, let $H_p\left[P\left(y\right|\cdot)\right]$ denote the identification region of the distribution $P\left(y\right|\cdot)$ given the observable data and the value $p \equiv P\left(z=1\right|d=0)$.

## A.1 Proposition 1

**Proof.** Denote by $\Psi$ the set of all probability distributions on $Y$; then by Proposition 1 in Horowitz and Manski (1995):

$$\begin{array}{l} P\left[y\right|d=0,z=1] \in H_p\left[P\left[y\right|d=0,z=1]\right] \\ H_p\left[P\left[y\right|d=0,z=1]\right] \equiv \Psi \cap \left\{ \frac{Q-(1-p)\psi}{p}, \psi \in \Psi \right\} \end{array} \tag{29}$$

Horowitz and Manski (1995) show that $L_p$ and $U_p$ are respectively stochastically dominated and stochastically dominating every other element of $H_p\left[P\left[y\right|d=0,z=1]\right]$. The result in (5) follows because the mean is a parameter that respects stochastic dominance. When the treatments are observable, the sharp bound on the average outcome under a mandatory policy is well known:

$$\begin{aligned} E\left[y\right|d=1,z=1]P\left[z=1\right|d=1] + K_0 P\left[z=0\right|d=1] &\leq P\left[y\left(1\right)=1\right|d=1] \\ &\leq P\left[y=1\right|d=1,z=1]P\left[z=1\right|d=1] + K_1 P\left[z=0\right|d=1] \end{aligned}$$

Using the Law of Iterated Expectations the result in (6) follows. To get the result in (**??**), notice that in absence of information on the distribution of the missing treatments, all that we can learn on $E\left[y\left(1\right)\right|d=0]$ is that it ranges in $[K_0, K_1]$. Hence, the result follows.

■

## A.2 Proposition 2

**Proof.** To find the sharp upper and lower bounds described in Proposition 2, observe that in any $UB$, $E_{10}=K_1$ and $E_{01}=K_0$, while in any $LB$, $E_{10}=K_0$ and $E_{01}=K_1$. Moreover, Proposition 4 in Horowitz and Manski (1995) implies that

$$\left(L_p, U_{1-p}\right) \in H_{p,(1-p)}\left[P\left[y\right|d=0,z=1], P\left[y\right|d=0,z=0]\right].$$

Note that $L_p$ is stochastically dominated by every member of $H_p\left[P\left[y\right|d=0,z=1]\right]$, while $U_{1-p}$ stochastically dominates every member of $H_{1-p}\left[P\left[y\right|d=0,z=0]\right]$. Hence the lower bound on $E_{11}p - E_{00}\left(1-p\right)$ will be given by $p\int ydL_p - (1-p)\int ydU_{1-p}$. Similar considerations hold for the upper bound. The Law of Iterated Expectations and expression (9) imply (14). To get the result in (15), take $\inf\limits_{p\in[0,1]} LB_{CTE}^{d=0}$ and $\sup\limits_{p\in[0,1]} UB_{CTE}^{d=0}$. Then use the Law of Iterated Expectations to get the result in (16).

■

### A.3 Proposition 3

**Proof.** Under the monotone treatment response assumption, we have: $E_{10} \in [K_0, E_{00}]$ and $E_{01} \in [E_{11}, K_1]$. The quantities that we want to bound are: $E[y(1)|\, d=0] = E_{11}p + E_{10}(1-p)$ and $E[y(0)=1|\, d=0] = E_{00}(1-p) + E_{01}p$. Since we know that $E[y|\, d=0] = E_{11}p + E_{00}(1-p)$, and that in any upper bound $E_{10}$ and $E_{01}$ are at their highest possible value, and in any lower bound they are at their lowest possible value, the result in (20) follows. Using the Law of Iterated Expectations and given (18)-(19), we get the results in (21)-(22). To get the lower bound in (23), subtract from the lower bound in (22) the upper bound in (21); to get the upper bound in (23), subtract from the upper bound in (22) the lower bound in (21).

■

# B Asymptotic Distribution of the Estimated Lower and Upper Bounds

As a first step, I introduce sample analog estimators of the population parameters appearing in the population bounds; for ease of notation, I continue to omit covariates in all that follows. Let $\hat{\theta}_1 \equiv \widehat{E}(y) = \frac{1}{N}\sum_{i=1}^{N} y_i$ denote the sample analog of $E(y)$, $\hat{\theta}_2 \equiv \widehat{\Pr}(d=1) = \frac{1}{N}\sum_{i=1}^{N} d_i$ denote the sample analog of $\Pr(d=1)$, and

$$\hat{\theta}_3 \equiv \widehat{\Pr}(z=1|\, d=1) = \frac{\frac{1}{N}\sum_{i=1}^{N} z_i d_i}{\frac{1}{N}\sum_{i=1}^{N} d_i}$$

denote the sample analog of $\Pr(z=1|\, d=1)$. Using similar notation, let $E(y|\, d=1)$ and $E(y|\, d=0)$ be estimated by

$$\hat{\theta}_4 \equiv \widehat{E}(y|\, d=1) = \frac{\frac{1}{N}\sum_{i=1}^{N} y_i d_i}{\frac{1}{N}\sum_{i=1}^{N} d_i} \quad \text{and} \quad \hat{\theta}_5 \equiv \widehat{E}(y|\, d=0) = \frac{\frac{1}{N}\sum_{i=1}^{N} y_i (1-d_i)}{\frac{1}{N}\sum_{i=1}^{N} (1-d_i)},$$

and $E(y|\, z=1, d=1)$ and $E(y|\, z=0, d=1)$ be estimated by

$$\hat{\theta}_6 \equiv \widehat{E}(y|\, z=1, d=1) = \frac{\frac{1}{N}\sum_{i=1}^{N} y_i z_i d_i}{\frac{1}{N}\sum_{i=1}^{N} z_i d_i} \quad \text{and} \quad \hat{\theta}_7 \equiv \widehat{E}(y|\, z=0, d=1) = \frac{\frac{1}{N}\sum_{i=1}^{N} y_i (1-z_i) d_i}{\frac{1}{N}\sum_{i=1}^{N} (1-z_i) d_i}.$$

Let $\hat{p}$ denote a root$-N$ consistent and asymptotically normal estimator of $p$. Finally let $\tau_L \equiv \int y dL_p$, $\tau_U \equiv \int y dU_p$, $\tau_{L,1-p} \equiv \int y dL_{1-p}$, and $\tau_{U,1-p} \equiv \int y dU_{1-p}$ be estimated by

$$\hat{\theta}_8 \equiv \hat{\tau}_L = \frac{\frac{1}{N\hat{p}}\sum_{i=1}^{N} y_i (1-d_i) I(y_i \leq \hat{r}(\hat{p}))}{\frac{1}{N}\sum_{i=1}^{N} (1-d_i)}$$

$$\hat{\theta}_9 \equiv \hat{\tau}_U = \frac{\frac{1}{N\hat{p}}\sum_{i=1}^{N} y_i (1-d_i) I(y_i \geq \hat{r}(1-\hat{p}))}{\frac{1}{N}\sum_{i=1}^{N} (1-d_i)}$$

$$\hat{\theta}_{10} \equiv \hat{\tau}_{L,1-p} = \frac{\frac{1}{N(1-\hat{p})} \sum_{i=1}^{N} y_i (1-d_i) I\left(y_i \leq \hat{r}\left(1-\hat{p}\right)\right)}{\frac{1}{N} \sum_{i=1}^{N} (1-d_i)}$$

$$\hat{\theta}_{11} \equiv \hat{\tau}_{U,1-p} = \frac{\frac{1}{N(1-\hat{p})} \sum_{i=1}^{N} y_i (1-d_i) I\left(y_i \geq \hat{r}\left(\hat{p}\right)\right)}{\frac{1}{N} \sum_{i=1}^{N} (1-d_i)}$$

where $\hat{r}(\theta)$ is the $\theta-$quantile of the empirical distribution function of $y|\, d=0$. Under a proper set of assumptions on the data generating process listed below, the above estimators are root$-N$ consistent and asymptotically normal.[10] Such assumptions are:

**Assumption 1:** A random sample $\{y_i, d_i, z_i d_i\}$, $i = 1, \ldots, N$ is available, such that $Var(y) < \infty$, $0 < \Pr(d=1) < 1$, $0 < \Pr(z=1|\, d=1) < 1$.

**Assumption 2:** If an auxiliary sample of size $n = \frac{N}{\kappa}$, $0 < \kappa < \infty$, is available, such that $p \equiv \Pr(z=1|\, d=0)$ can be consistently estimated using this sample, the auxiliary sample is independent from the sample used to estimate the distribution of $y|\, d=0$. Let $\hat{p}$ denote the estimator of $p$. Then $\hat{p}$ satisfies: $\hat{p} = 0\,(1)$ if $p = 0\,(1)$, and $\sqrt{n}(\hat{p}-p) \xrightarrow{d} N(0, V_p)$ otherwise, so that $\sqrt{N}(\hat{p}-p) \xrightarrow{d} N(0, \kappa V_p)$.

**Assumption 3:** $Q$ is continuously differentiable in neighborhoods of $r(p)$ and $r(1-p)$, and $Q'(r(p)) > 0$ and $Q'(r(1-p)) > 0$.

The first assumption simply states that the researcher observes a random sample from the population of interest. The second assumption requires that if the researcher can estimate $\hat{p}$ from an auxiliary sample, then this auxiliary sample has to be independent from that used to estimate the distribution of $y|\, d=0$. This assumption is necessary to obtain the asymptotic distribution of $\hat{\tau}_L$ and $\hat{\tau}_U$, which are randomly trimmed means. The random trimming proportion is given by $\hat{p}$, and the distribution to be trimmed is the empirical distribution function of $y|\, d=0$. For example, if one assumes that the treatments are missing at random, then $\hat{p} = \widehat{\Pr}(z=1|\, d=1)$, and the independence assumption is satisfied. Assumption 3 is a regularity condition needed to obtain the asymptotic distribution of $\hat{\tau}_L$ and $\hat{\tau}_U$ (Shorack and Wellner (1986), Chapter 19).

For reasons of brevity, rather than obtaining the confidence sets for each of the identification regions of Sections 4-5, I derive below the joint asymptotic distribution of all estimators listed above. Given this asymptotic distribution, one can easily obtain the joint distribution of the estimators of each of the lower and upper bounds in Sections 4-5 by using the delta method, and then obtain confidence sets as detailed in Imbens and Manski (2004) and Chernozhukov, Hong and Tamer (2004, Section 3.3). Alternatively, one can use the bootstrap to estimate the variance-covariance matrix of the lower and upper bounds of interest.

[10] Root$-N$ consistency and asymptotic normality of the estimators of population means follows from standard arguments. For a discussion of the asymptotic properties of randomly trimmed means, see Shorack and Wellner (1986), Chapter 19. Related results are derived by Horowitz and Manski (1997).

Let $\hat{\boldsymbol{\theta}} = \left[\hat{\theta}_1, \ldots, \hat{\theta}_{11}\right]'$, with $\boldsymbol{\theta}$ its population counterpart. Then, under Assumptions 1-3 the following result holds:

$$\sqrt{N}\begin{pmatrix} \hat{\boldsymbol{\theta}} - \boldsymbol{\theta} \\ \hat{p} - p \end{pmatrix} \xrightarrow{d} N\left(\mathbf{0}, \Sigma\right),$$

where, denoting the elements of $\Sigma$ by $\sigma_{ij}$, $\sigma_{11} = Var\left(y\right)$, $\sigma_{22} = \Pr\left(d=1\right)\left[1-\Pr\left(d=1\right)\right]$, $\sigma_{33} = \frac{\Pr(z=1|d=1)[1-\Pr(z=1|d=1)]}{\Pr(d=1)}$, $\sigma_{44} = \frac{Var(y|d=1)}{\Pr(d=1)}$, $\sigma_{55} = \frac{Var(y|d=0)}{\Pr(d=0)}$, $\sigma_{66} = \frac{Var(y|z=1,d=1)}{\Pr(z=1,d=1)}$, $\sigma_{77} = \frac{Var(y|z=0,d=1)}{\Pr(z=0,d=1)}$, $\sigma_{12,12} = \kappa V_p$. Shorack and Wellner (1986, Chapter 19, Theorem 1) show that

$$\sqrt{N}\left(\hat{\tau}_L - \tau_L\right) \underset{a}{=} -\frac{1}{p}\left\{\int_{K_0}^{r(p)} \mathbf{U} dr - \left(r\left(p\right) - \tau_L\right)\sqrt{N}\left(\hat{p} - p\right)\right\}$$

where $\mathbf{U}$ is the empirical process of the Uniform order statistics corresponding to the order statistics of $y_1, \ldots, y_N | \, d = 0$. Similar results hold for $\hat{\tau}_U$, $\hat{\tau}_{L,1-p}$, and $\hat{\tau}_{U,1-p}$. Given the independence of $\hat{p}$ and $\hat{Q}$, where $\hat{Q}$ denotes the empirical distribution function of $y | \, d = 0$, it follows that

$$\begin{aligned}
\sigma_{88} &= \frac{\left[p(1-p)[r(p)]^2 + \int_{K_0}^{r(p)} y^2 dQ - p\tau_L^2 + p(1-p)\tau_L^2 - 2p(1-p)\tau_L r(p)\right] + \kappa V_p (r(p) - \tau_L)^2}{p^2}, \\
\sigma_{99} &= \frac{\left[p(1-p)[r(1-p)]^2 + \int_{r(1-p)}^{K_1} y^2 dQ - p\tau_U^2 + p(1-p)\tau_U^2 - 2p(1-p)\tau_U r(1-p)\right] + \kappa V_p (\tau_U - r(1-p))^2}{p^2}, \\
\sigma_{10,10} &= \frac{\left[p(1-p)[r(1-p)]^2 + \int_{K_0}^{r(1-p)} y^2 dQ - (1-p)\tau_{L,1-p}^2 + p(1-p)\tau_{L,1-p}^2 - 2p(1-p)\tau_{L,1-p} r(1-p)\right] + \kappa V_p \left(r(1-p) - \tau_{L,1-p}\right)^2}{(1-p)^2} \\
\sigma_{11,11} &= \frac{\left[p(1-p)[r(p)]^2 + \int_{r(p)}^{K_1} y^2 dQ - (1-p)\tau_{U,1-p}^2 + p(1-p)\tau_{U,1-p}^2 - 2p(1-p)\tau_{U,1-p} r(p)\right] + \kappa V_p \left(\tau_{U,1-p} - r(p)\right)^2}{(1-p)^2}.
\end{aligned}$$

Regarding the covariances, those involving products of sample averages can be derived using standard results, and are available from the author upon request. Here I focus on providing the covariances involving $\hat{\tau}_L$, for the case that $\hat{p} = \widehat{\Pr}\left(z = 1 | \, d = 1\right)$; those involving $\hat{\tau}_U$ can be obtained analogously. In particular,

$$\begin{aligned}
\sigma_{18} &= \tfrac{r(p)-\tau_L}{p}\tfrac{[E(y|z=1,d=1) - E(y|d=1)]\Pr(z=1,d=1)}{\Pr(d=1)} + \tfrac{\Pr(d=0)}{p}\int_0^1 \int_0^p \left(\min\left(s,t\right) - st\right) dr\left(s\right) dr\left(t\right), \\
\sigma_{28} &= 0, \\
\sigma_{38} &= \tfrac{r(p)-\tau_L}{p}\tfrac{\Pr(z=1|d=1)[1-\Pr(z=1|d=1)]}{\Pr(d=1)}, \\
\sigma_{48} &= \tfrac{r(p)-\tau_L}{p}\tfrac{[E(y|z=1,d=1) - E(y|d=1)]\Pr(z=1|d=1)}{\Pr(d=1)}, \\
\sigma_{58} &= \tfrac{1}{p}\int_0^1 \int_0^p \left(\min\left(s,t\right) - st\right) dr\left(s\right) dr\left(t\right), \\
\sigma_{68} &= 0, \\
\sigma_{78} &= 0, \\
\sigma_{98} &= \tfrac{1}{p^2}\int_{1-p}^1 \int_0^p \left(\min\left(s,t\right) - st\right) dr\left(s\right) dr\left(t\right) - \tfrac{\tau_U - r(1-p)}{p}\tfrac{r(p)-\tau_L}{p}\kappa V_p, \\
\sigma_{10,8} &= \tfrac{1}{p(1-p)}\int_0^{1-p} \int_0^p \left(\min\left(s,t\right) - st\right) dr\left(s\right) dr\left(t\right) - \tfrac{r(1-p)-\tau_{L,1-p}}{1-p}\tfrac{r(p)-\tau_L}{p}\kappa V_p, \\
\sigma_{11,8} &= \tfrac{1}{p(1-p)}\int_p^1 \int_0^p \left(\min\left(s,t\right) - st\right) dr\left(s\right) dr\left(t\right) + \tfrac{\tau_{U,1-p} - r(p)}{1-p}\tfrac{r(p)-\tau_L}{p}\kappa V_p, \\
\sigma_{12,8} &= \kappa V_p \tfrac{r(p)-\tau_L}{p}.
\end{aligned}$$

Given these results, one can show, for example, that the joint asymptotic distribution of the sample analogs of the lower and upper bounds on $E\left[y\left(1\right)\right|d=0]$ as in equation (5), given by $\vartheta_L = p\int ydL_p + (1-p)K_0$, $\vartheta_U = p\int ydU_p + (1-p)K_1$, is as follows:

$$\sqrt{N}\left(\begin{array}{c}\hat{\vartheta}_L - \vartheta_L \\ \hat{\vartheta}_U - \vartheta_U\end{array}\right) \xrightarrow{d} \left(\begin{array}{c}W_1 \\ W_2\end{array}\right) \sim N\left(\mathbf{0}, \Omega\right),$$

where the elements of $\Omega$, denoted $\omega_{ij}$, are given by

$$\omega_{11} = \left[p\left(1-p\right)\left[r\left(p\right)\right]^2 + \int_{K_0}^{r(p)} y^2 dQ - p\tau_L^2 + p\left(1-p\right)\tau_L^2 - 2p\left(1-p\right)\tau_L r\left(p\right)\right] + \kappa V_p\left(r\left(p\right) - K_0\right)^2,$$

$$\begin{aligned}\omega_{22} &= \left[p\left(1-p\right)\left[r\left(1-p\right)\right]^2 + \int_{r(1-p)}^{K_1} y^2 dQ - p\tau_U^2 + p\left(1-p\right)\tau_U^2 - 2p\left(1-p\right)\tau_U r\left(1-p\right)\right] \\ &\quad + \kappa V_p\left(r\left(1-p\right) - K_1\right)^2,\end{aligned}$$

$$\omega_{12} = \int_{1-p}^{1}\int_{0}^{p}\left(\min\left(s,t\right) - st\right)dr\left(s\right)dr\left(t\right) - \left(K_1 - r\left(1-p\right)\right)\left(r\left(p\right) - K_0\right)\kappa V_p,$$

Confidence intervals $CI^{\vartheta^\star}$ that asymptotically cover $\vartheta^\star \equiv E\left[y\left(1\right)\right|d=0]$ with probability at least $\left(1-\alpha\right)$, as in Imbens and Manski (2004), can be obtained by calculating the value of $c_\alpha$ such that

$$\Phi\left(c_\alpha + \sqrt{N}\frac{\hat{\vartheta}_U - \hat{\vartheta}_L}{\max\left(\sqrt{\hat{\omega}_{11}}, \sqrt{\hat{\omega}_{22}}\right)}\right) - \Phi\left(-c_\alpha\right) = 1-\alpha,$$

and setting

$$CI^{\vartheta^\star} = \left[\hat{\vartheta}_L - \frac{c_\alpha\sqrt{\hat{\omega}_{11}}}{\sqrt{N}}, \hat{\vartheta}_U + \frac{c_\alpha\sqrt{\hat{\omega}_{22}}}{\sqrt{N}}\right]$$

Confidence intervals $CI^{[\vartheta_L,\vartheta_U]}$ that asymptotically cover $[\vartheta_L, \vartheta_U]$ with probability $\left(1-\alpha\right)$, as in Chernozhukov, Hong, and Tamer (2004), can be obtained by calculating the value of $c_\alpha$ such that

$$\Pr\left[\max\left(\left(W_1\right)_+^2, \left(W_2\right)_-^2\right) \leq c_\alpha\right] = 1-\alpha,$$

where $\left(u\right)_+^2 = \left(u\right)^2 \cdot I\left[u>0\right]$ and $\left(u\right)_-^2 = \left(u\right)^2 \cdot I\left[u<0\right]$, and setting

$$CI^{\vartheta^\star} = \left[\hat{\vartheta}_L - \sqrt{\frac{c_\alpha}{N}}, \hat{\vartheta}_U + \sqrt{\frac{c_\alpha}{N}}\right].$$

## References


Angrist, J. D., G. W. Imbens, and D. B. Rubin (1996): "Identification of Causal Effects Using Instrumental Variables," *Journal of the American Statistical Association*, 91(434), 444–455.

Balke, A., and J. Pearl (1994): "Counterfactual Probabilities: Computational Methods, Bounds and Applications," *Proceedings of the Conference on Uncertainty in Artificial Intelligence (UAI-94)*, pp. 46–54.

——— (1995): "Universal Formulas for Treatment Effects from Noncompliance Data," in *Lifetime Data: Models in Reliability and Survival Analysis*, ed. by N. P. Jewell, A. C. Kimber, M. L. Lee, and G. A. Whitmore, pp. 39–43. Kluwer Academic Publishers.

——— (1997): "Bounds on Treatment Effects from Studies with Imperfect Compliance," *Journal of the American Statistical Association*, 92(439), 1171–1176.

BLS (1999): "NLSY79 User's Guide," *National Longitudinal Surveys*, U.S. Department of Labor, Bureau of Labor Statistics.

Buchmueller, T., and S. Zuvekas (1998): "Drug Use, Drug Abuse, and Labor Market Outcomes," *Health Economics*, 7, 229–45.

Chavez-Martin-Del-Campo, J. C. (2005): "Partial Identification of Poverty Measures with Contaminated and Corrupted Data," Discussion paper, Cornell University.

Chernozhukov, V., H. Hong, and E. Tamer (2004): "Parameter Set Inference in a Class of Econometric Models," Discussion paper, Northwestern University.

Gill, A., and R. Michaels (1992): "Does Drug Use Lower Wages?," *Industrial and Labor Relations Review*, 45(3), 419–434.

Gleason, P. M., J. R. Veum, and M. R. Pergamit (1991): "Drug and Alcohol Use at Work: a Survey of Young Workers," *Monthly Labor Review*, pp. 3–7.

Harwood, H., D. Fountain, and G. Livermore (1998): "The Economic Costs of Alcohol and Drug Abuse in the United States, 1992," *National Institute on Drug Abuse (NIDA) and National Institute on Alcohol Abuse and Alcoholism (NIAAA)*.

Horowitz, J. L., and C. F. Manski (1995): "Identification and Robustness with Contaminated and Corrupted Data," *Econometrica*, 63(2), 281–302.

——— (1997): "What Can Be Learned About Population Parameters When the Data are Contaminated," in *Handbook of Statistics, Vol. 15*, ed. by G. S. Maddala, and C. R. Rao, pp. 439–466, North Holland. Elsevier Science B. V.

Horowitz, J. L., and C. F. Manski (1998): "Censoring of Outcomes and Regressors Due to Survey Nonresponse: Identification and Estimation Using Weights and Imputations," *Journal of Econometrics*, 84, 37–58.

Horowitz, J. L., and C. F. Manski (2000): "Nonparametric Analysis of Randomized Experiments With Missing Covariate and Outcome Data," *Journal of the American Statistical Association*, 95(449), 77–84.

Hotz, V. J., C. H. Mullin, and S. G. Sanders (1997): "Bounding Causal Effects Using Data from a Contaminated Natural Experiment: Analyzing the Effects of Teenage Childbearing," *Review of Economic Studies*, 64, 575–603.

Imbens, G. W., and C. F. Manski (2004): "Confidence Intervals for Partially Identified Parameters," *Econometrica*, 72(6), 1845–1857.

Imbens, G. W., and W. A. Pizer (1999): "The Analysis of Randomized Experiments with Missing Data," *Discussion Paper 00-19, Resources for the Future*, http://www.rff.org.

Kaestner, R. (1991): "The Effect of Illicit Drug Use on the Wages of Young Adults," *Journal of Labor Economics*, 9(4), 381–412.

——— (1994): "The Effect of Illicit Drug Use on the Labor Supply of Young Adults," *Journal of Human Resources*, 29(1), 123–136.

——— (1998): "Illicit Drug Use and Labor Market Outcomes: A Review of Economic Theory and Its Empirical Applications," *Journal of Drug Issues*, 28(3), 663–680.

Kaestner, R., and M. Grossman (1995): "Wages, Workers' Compensation Benefits, and Drug Use: Indirect Evidence of the Effect of Drugs on Workplace Accidents," *The American Economic Review, Papers and Proceedings*, 85(2), 55–60.

Kandel, D. B., and M. Davies (1990): "Labor Force Experiences of a National Sample of Young Adult Men: The Role of Drug Involvement," *Youth and Society*, 21(4), 411–445.

Kandel, D. B., and K. Yamaguchi (1987): "Job Mobility and Drug Use: An Event History Analysis," *American Journal of Sociology*, 92(4), 836–878.

Kreider, B., and S. C. Hill (2005): "Partial Identification of Treatment Effects in the Presence of Unobserved Treatments: The Case of Universal Health Insurance," Discussion paper, Iowa State University.

Little, R. J. A. (1992): "Regression with Missing X's: A Review," *Journal of the American Statistical Association*, 87, 1227–1237.

Little, R. J. A., and D. B. Rubin (1987): *Statistical Analysis with Missing Data*. John Wiley and Sons.

Manski, C. F. (1995): *Identification Problems in the Social Sciences*. Harvard University Press.

——— (1997): "Monotone Treatment Response," *Econometrica*, 65(6), 1311–1334.

Manski, C. F., and J. V. Pepper (2000): "Monotone Instrumental Variables: with an Application to the Returns on Schooling," *Econometrica*, 68(4), 997–1010.

Mensch, B. S., and D. B. Kandel (1998): "Underreporting of Substance Use in a National Longitudinal Youth Cohort," *Public Opinion Quarterly*, 52, 100–124.

Molinari, F. (2003): "Contaminated, Corrupted, and Missing Data," Ph.D. thesis, Northwestern University.

——— (2005): "Partial Identification of Probability Distributions with Misclassified Data," Discussion paper, Cornell University.

Mullahy, J., and J. L. Sindelar (1995): "Employment, Unemployment, and Problem Drinking," NBER Working Paper No. 5123.

Neyman, J. S. (1923): "On the Application of Probability Theory to Agricaltural Experiments. Essay on Principles. Section," *Roczniki Nauk Rolniczych.*, 10, 1–51, Reprinted (Translated and Edited) in *Statistical Science, 1990, Vol. 5, No. 4, 465-480.*

Pepper, J. V. *(2001): "How Do Response Problems Affect Survey Measurement of Trends in Drug Use?," in* Informing America's Policy on Illegal Drugs: What We Don't Know Keeps Hurting Us*, ed. by C. F. Manski, J. V. Pepper, and C. Petrie, pp. 321–348, Washington D.C. National Academy Press.*

Register, C., and D. Williams *(1992): "Labor Market Effects of Marijuana and Cocaine Use Among Young Men,"* Industrial and Labor Relations Review*, 45(3), 434–448.*

ROBINS, J. M. *(1989): "The Analysis of Randomized and Nonrandomized AIDS Treatment Trials Using a New Approach to Causal Inference in Longitudinal Studies," in* Health Service Research Methodology: A Focus on AIDS*, ed. by L. Sechrest, H. Freeman, and A. Mulley, pp. 113–159, Washington D.C. U.S. Public Health Service, National Center for Health Services Research.*

——— *(1997): "Non-Response Models for the Analysis of Non-Monotone Non-Ignorable Missing Data,"* Statistics in Medicine*, 16, 21–37.*

ROBINS, J. M., A. ROTNIZKY, AND L. ZHAO *(1994): "Estimation of Regression Coefficients When Some Regressors Are Not Always Observed,"* Journal of the American Statistical Association*, 89, 846–866.*

SERFLING *(1984):* Limit Theorems for Mathematical Statistics*. John Wiley and Sons.*

SHORACK, G. R. *(1997): "Uniform CLT, WLLN, LIL and Bootstrapping in a Data Analytic Approach to Trimmed L-Statistics,"* Journal of Statistical Planning and Inference*, 60, 1–44.*

SHORACK, G. R., AND J. A. WELLNER *(1986):* Empirical Processes with Applications to Statistics*. John Wiley and Sons, New York.*

WANG, C., S. WANG, L. ZHAO, AND S. OU *(1997): "Weighted Semiparametric Estimation in Regression Analysis with Missing Covariate Data,"* Journal of the American Statistical Association*, 92, 512–525.*