Front-door Versus Back-door Adjustment with Unmeasured Confounding: Bias Formulas for Front-door and Hybrid Adjustments*

Adam Glynn† Konstantin Kashin‡

aglynn@fas.harvard.edu kkashin@fas.harvard.edu

February 14, 2014

Abstract

In this paper, we develop bias formulas for front-door and front-door/back-door hybrid estimators that utilize information from post-treatment variables under general patterns of measured and unmeasured confounding. These bias formulas allow for sensitivity analysis, and also allow for comparisons to the bias resulting from standard pre-treatment covariate adjustments such as matching or regression adjustments (also known as back-door adjustments). We also present these bias comparisons in two special cases: nonrandomized program evaluations with one-sided noncompliance and linear structural equation models. These comparisons demonstrate that there are broad classes of applications for which the front-door or hybrid adjustments will be preferred to back-door adjustments. These comparisons also have surprising implications for the design of observational studies. First, the measurement of auxiliary post-treatment variables may be as important as the measurement of some pre-treatment covariates, even in the assessment of total effects. Second, in some applications it will not be necessary to collect any information on control units. We illustrate these points with an application to the National JTPA (Job Training Partnership Act) Study.

Keywords: *Causal inference, post-treatment, program evaluation, sensitivity analysis.*

*Winner of the 2013 Gosnell Prize for Excellence in Political Methodology. We would like to acknowledge Jeffrey Smith and Petra Todd for generously sharing their data from the National JTPA Study. We also thank Nahomi Ichino, Kosuke Imai, Gary King, Judea Pearl, Kevin Quinn, and Teppei Yamamoto for comments and suggestions.

†Department of Government and Institute for Quantitative Social Science, Harvard University, 1737 Cambridge Street, Cambridge, MA 02138 (http://scholar.harvard.edu/aglynn).

‡Department of Government and Institute for Quantitative Social Science, Harvard University, 1737 Cambridge Street, Cambridge, MA 02138 (http://konstantinkashin.com).

1 Introduction

The front-door criterion and adjustment formula (Pearl, 1995) and its extensions provide a means for nonparametric identification of treatment effects via the pathways or mechanisms by which treatment affects the outcome.[1] Importantly, the front-door criterion can hold even in the presence of unmeasured common causes of the treatment and the outcome. However, the front-door approach has seen relatively little use (VanderWeele, 2009) due to concerns that the assumptions required for point identification are "exceptional" (Cox and Wermuth, 1995; Imbens and Rubin, 1995).[2]

In this paper, we consider the applicability of the front-door adjustment in situations where the front-door criterion does not hold exactly, by providing formulas for the large sample bias of front-door adjustments for both Average Treatment Effects on the Treated (ATT) and Average Treatment Effects (ATE). These formulas allow for sensitivity analysis and are derived without using potential outcomes beyond those that are used for the definition of ATT and ATE. Therefore, they do not require causal effects to be well defined for variables other than the treatment. This allows for direct comparisons to the bias formulas of VanderWeele and Arah (2011) for standard covariate adjustments (sometimes known as back-door adjustment).

To provide intuition, we also present these bias comparisons in two special cases: the estimation of ATT for nonrandomized program evaluations with one-sided noncompliance, and the estimation of ATE using linear structural equation models. These comparisons demonstrate that there are

[1] Extensions of the front-door criterion have highlighted more complicated graph structures under which it is possible to obtain point identification of total effects (Kuroki and Miyakawa, 1999; Tian and Pearl, 2002a,b; Shpitser and Pearl, 2006; Chalak and White, 2011).

[2] One exception is Winship and Harding (2008) which outlines how the front-door criterion can aid in the identification of age-period-cohort models. Additionally, there are several papers that use post-treatment variables to gain some type of information about total effects. Cox (1960) and Ramsahai (2012) examine when post-treatment variables can improve the efficiency of total effects estimates. VanderWeele (2008) and VanderWeele and Robins (2009) show that post-treatment variables can help identify the direction of bias in point estimates of total effects. Joffe (2001) and Glynn and Quinn (2011) both use post-treatment variables to calculate bounds for total effects while Kaufman et al. (2009) provides a variety of bounds, some of which involve measuring post-treatment variables, using linear programming via the OPTIMIZE program (Balke, 1995).

broad classes of applications for which the front-door or hybrid adjustments will be preferred to the back-door adjustments.

To further illustrate the applicability of the front-door approach, we demonstrate the use of the technique for nonrandomized program evaluations with an application to the National JTPA (Job Training Partnership Act) Study. We show that by using information on enrollment in the program in addition to some standard pre-treatment covariates, the front-door adjustment provides estimates that are closer to the experimental benchmark than standard covariate adjustments. Interestingly, the front-door adjustment in this application does not utilize data on control units.

The paper is organized as follows. Section 2 presents the bias formulas for the front-door approach to ATT and compares this to the bias from covariate adjustments, both within the general framework and for nonrandomized program evaluations with one-sided noncompliance. Section 3 presents an application of these methods to the National JTPA (Job Training Partnership Act) Study. Section 4 concludes. Because the presentation for ATE is somewhat parallel and redundant to the presentation for ATT, we provide the bias formulas and comparisons, both within the general framework and for linear structural equation models, in the supplementary material.

2 The Front-Door Approach for ATT

For an outcome Y and a treatment/action A, we define the potential outcome under a generic treatment as $Y(a_1)$ and the potential outcome under control as $Y(a_0)$. While the presentation of the front-door approach in Pearl (1995, 2000, 2009) focuses on ATE ($E[Y(a_1)] - E[Y(a_0)]$), in many applications ATT ($E[Y(a_1)|a_1] - E[Y(a_0)|a_1]$) is the question of interest. See Supplement A for an extended discussion of the front-door adjustment for ATE.

We assume consistency such that $\mu_{1|a_1} = E[Y(a_1)|a_1]$ equals the observable quantity $E[Y|a_1]$. We also assume that $E[Y(a_0)|a_1]$ is identifiable by conditioning on observed covariates X and unobserved covariates U. For simplicity in presentation we assume that X and U are discrete, such

that

$$\mu_{0|a_1} = E[Y(a_0)|a_1] = \sum_x \sum_u E[Y|a_0, x, u] \cdot P(u|x, a_1) \cdot P(x|a_1), \tag{1}$$

but continuous variables can be easily accommodated. Note that the form of this equation represents a non-trivial assumption, because it requires that positivity holds such that the conditional distributions are well defined.

The front-door adjustment for a set of measured post-treatment variables M can be written as the following:

$$\mu^{fd}_{0|a_1} = \sum_x \sum_m P(m|a_0, x) \cdot E[Y|a_1, m, x] \cdot P(x|a_1) \tag{2}$$

The bias in the front-door estimate of $E[Y(a_0)|a_1]$ is the following (see 15 for a proof):

$$\begin{aligned} B^{fd}_{0|a_1} = &\sum_x P(x|a_1) \sum_m \sum_u P(m|a_0, x) \cdot E[Y|a_1, m, x, u] \cdot P(u|a_1, m, x) \\ &- \sum_x P(x|a_1) \sum_m \sum_u P(m|a_0, x, u) \cdot E[Y|a_0, m, x, u] \cdot P(u|a_1, x) \end{aligned} \tag{3}$$

Note that the bias will be zero when Y is mean independent of A conditional on U, M, and X (i.e., $E[Y|a_1, m, x, u] = E[Y|a_0, m, x, u]$) and U is independent of M conditional on X and a_0 or a_1 (i.e., $P(m|a_0, x) = P(m|a_0, x, u)$ and $P(u|a_1, m, x) = P(u|a_1, x)$). Hence, as shown in Pearl (1995) for ATE, it is possible for the front-door approach to provide an unbiased estimator when there is an unmeasured confounder.

The back-door estimator of $E[Y(a_0)|a_1]$ and the bias of this estimator can be written as the following (see 16 for a proof):

$$\mu^{bd}_{0|a_1} = \sum_x E[Y|a_0, x] \cdot P(x|a_1) \tag{4}$$

$$B^{bd}_{0|a_1} = \sum_x P(x|a_1) \sum_u E[Y|a_0, x, u][P(u|a_0, x) - P(u|a_1, x)] \tag{5}$$

This is very similar to the formula presented in VanderWeele and Arah (2011). Since consistency implies that $E[Y(a_1)|a_1] = E[Y|a_1]$, the front-door ATT bias is $B^{fd}_{ATT} = -B^{fd}_{0|a_1}$ and the back-door ATT bias is $B^{bd}_{ATT} = -B^{bd}_{0|a_1}$. Hence, the front-door ATT bias can be smaller than the back-door ATT bias even when the aforementioned front-door independence conditions do not hold exactly. It is also possible to form hybrid estimators that utilize the front-door estimate for some values of X and the back-door estimate for other values of X. Finally, we note that these are direct comparisons in the sense that we did not define additional potential outcomes in order to derive the front-door result (i.e., we are agnostic as to whether M is a set of well defined treatment variables). In fact, we do not strictly need M to be causally prior to Y in order for these formulas to be valid. In this sense, our conditions for identification are more general than those presented in Pearl (1995).

2.1 Special Case: Nonrandomized Program Evaluations with One-Sided Noncompliance

In order to develop some intuition about ATT estimators, we next consider the special case of nonrandomized program evaluations with one-sided noncompliance. Following a robust literature in econometrics on social program evaluation, we define the program impact as the ATT where the active treatment (a_1) is assignment into a program (Heckman et al., 1998).[3]

Consider the bias in the front-door estimator for ATT when M indicates whether the active treatment (a_1) was actually received and there is one-sided noncompliance such that $P(M = 0|a_0, x) =$

[3]There is some ambiguity regarding the use of the term ATT. Some authors refer to the parameter of interest as ATT, however, once noncompliance is emphasized, other authors might refer to the parameter as the Intent to Treat Effect on the Intended (ITI). This will be discussed further below.

$P(M = 0|a_0, x, u) = 1$ for all x, u. In this case, the front-door estimator reduces to the following:

$$\begin{aligned}\mu^{fd}_{ATT} &= \mu_{1|a_1} - \mu^{fd}_{0|a_1} \\ &= E[Y|a_1] - \sum_x \sum_m P(m|a_0, x) \cdot E[Y|a_1, m, x] \cdot P(x|a_1) \\ &= E[Y|a_1] - \sum_x \underbrace{E[Y|a_1, M = 0, x]}_{\text{treated non-compliers}} \cdot P(x|a_1)\end{aligned} \tag{6}$$

Compare this to the standard back-door estimator for ATT:

$$\begin{aligned}\mu^{bd}_{ATT} &= \mu_{1|a_1} - \mu^{bd}_{0|a_1} \\ &= E[Y|a_1] - \sum_x \underbrace{E[Y|a_0, x]}_{\text{controls}} \cdot P(x|a_1)\end{aligned} \tag{7}$$

Intuitively, standard back-door estimates implicitly or explicitly match units that were assigned treatment to similar units that were assigned control. Front-door estimates implicitly or explicitly match units that were assigned treatment to similar units that were assigned treatment but did not receive treatment. More specifically, those that were assigned treatment and received treatment are implicitly matched to similar units that were assigned treatment but did not receive treatment. Those that were assigned treatment and did not receive treatment (i.e., non-compliers) are implicitly matched to themselves.

These sorts of comparisons (comparing treated units to treated units) are non-standard, so it is helpful to consider the intuitive justification for the technique before presenting the formal statement of bias. By matching non-compliers to themselves, the technique effectively assumes that treatment assignment has no effect if treatment is not received (i.e., an exclusion restriction holds). By matching units that are assigned and receive treatment to non-compliers, the technique effectively also assumes that non-compliance is assigned as if it were random (conditional on covariates). Of course these assumptions will likely not hold, but there may be applications where these assumptions are preferable to the assumptions required for standard estimators. The idea of leveraging

non-compliers in this manner was briefly explored in Heckman et al. (1997), although it was not mentioned in the abstract or conclusion, and it was not discussed in connection to the front-door approach.

The use of non-compliers in this manner also introduces some ambiguity regarding ATT as the parameter of interest due to the difference between treatment assigned and treatment received. Some authors continue to refer to the assigned treatment as "the treatment", and $\mu_{1|a_1} - \mu_{0|a_1}$ as ATT, while other authors would refer to the received treatment as "the treatment", and $\mu_{1|a_1} - \mu_{0|a_1}$ would be more properly characterized as the Intent to Treat Effect on the Intended (ITI). For continuity, we will continue to refer to $\mu_{1|a_1} - \mu_{0|a_1}$ as the ATT. This is consistent with the parameter of interest in the econometrics literature utilizing JTPA data (Heckman et al., 1998, 1997; Heckman and Smith, 1999), and is a relevant parameter from the point of view of policymakers since "[it] is informative on how the availability of a program affects participant outcomes" (Heckman et al., 1999).[4]

The front-door and the back-door ATT bias under one-sided noncompliance can be written as the following (see Appendix A.2):

$$B^{fd}_{ATT} = \sum_x P(x|a_1) \sum_u E[Y|a_0, M=0, x, u][P(u|a_1, x) - P(u|a_1, x, M=0)] \tag{8}$$

$$- \sum_x P(x|a_1) \sum_u \{E[Y|a_1, M=0, x, u] - E[Y|a_0, M=0, x, u]\} P(u|a_1, M=0, x) \tag{9}$$

$$B^{bd}_{ATT} = \sum_x P(x|a_1) \sum_u E[Y|a_0, M=0, x, u][P(u|a_1, x) - P(u|a_0, x)] \tag{10}$$

In order to compare the bias coming from the front-door and back-door approaches, it is helpful to separately consider the two components of front-door bias represented by (8) and (9). If we

[4]For interest in the Intent to Treat Effect outside of the JTPA program, see for example Lee (2009) and Zhang et al. (2009).

assume that the (9) component is zero for illustrative purposes, then the bias comparison between these two approaches is a comparison between (8) and (10). In this scenario, when the unobserved covariates of the treated units are matched better by the unobserved covariates for noncompliant treated units than by the unobserved covariates for the control units, then the front-door estimate will produce less bias.

As we consider in the next section, (8) will sometimes be smaller than (10), so the key question will often be the magnitude of (9). Note that an exclusion restriction (that A only affects Y through M) will likely be necessary in order for (9) to be zero.[5] Unfortunately, an exclusion restriction is not sufficient because conditioning on M can induce dependence between A and Y. For example, it is possible that an unmeasured variable $v \notin U$ is a common cause of both M and Y. Furthermore, induced dependence will occur with unmeasured variables $v, w \notin U$ such that v is a common cause of M and U and w is a common cause of U and Y. Similarly, induced dependence would occur if v is a common cause of M and X and w is a common cause of X and Y.

Despite these complications, it will be possible in some circumstances to glean information from these equations that will be useful in a comparison of front-door and back-door results. For example, if we are willing to assume that (8), (9), and (10) all have the same sign, then we should have a preference upon observing front-door and back-door estimates. An example of this will be provided in the next section. Alternatively, we might believe that (8) and (9) have opposite signs, and therefore front-door bias will be smaller in magnitude than bias from the back-door approach in (10). As we discuss in the conclusion, prior beliefs along these lines might have implications for research design.

2.1.1 Comparative Sensitivity Analysis

Equations (8) and (9) form the basis for a sensitivity analysis of the front-door approach, and this sensitivity analysis can be compared fruitfully with a sensitivity analysis for the back-door approach

[5] It is possible that cancelations would make it unnecessary.

based on (10). In order to illustrate this, we start with the simplifying assumptions used in VanderWeele and Arah (2011), although as discussed in that article it is straightforward to relax these assumptions at the cost of complicating the analysis.

VanderWeele and Arah (2011) shows that when U is binary and when

$$\gamma = E[Y|U = 1, a_0, M = 0, x] - E[Y|U = 0, a_0, M = 0, x] \tag{11}$$

and

$$\delta = P(U = 1|a_1, x) - P(U = 1|a_0, x) \tag{12}$$

do not depend on x, then (10) can be written as $\gamma \cdot \delta$. In this case, δ can be interpreted as the imbalance on U across the treatment and control groups, and γ can be interpreted as a sort of controlled direct effect of U on Y for controls, although this "effect" need not be a well defined causal effect. If we additionally assume that

$$\varepsilon = P(U = 1|a_1, x) - P(U = 1|a_1, M = 0, x) \tag{13}$$

does not depend on x, then (8) can similarly be re-written as $\gamma \cdot \varepsilon$, where ε can be interpreted as the imbalance on U between the treated units and treated noncompliers (see Appendix A.3 for proof).

If we also make the simplifying assumption that the weighted average of direct "effects" of A,

$$\eta = \sum_u \{E[Y|u, a_1, M = 0, x] - E[Y|u, a_0, M = 0, x]\} P(u|a_1, M = 0, x), \tag{14}$$

does not depend on x, then we can write the front-door bias as $\gamma \cdot \varepsilon - \eta$. Therefore, once one specifies the sensitivity parameters for a back-door estimator (γ and δ), one need only specify two more parameters for a sensitivity analysis of the front-door estimator (ε and η). Furthermore, these formulas illustrate the efficacy of the front-door approach. If we believe that the absolute value of

the direct "effect" η is small, and that the ε is smaller in absolute value than δ, then we will likely prefer the front-door approach to the back-door approach. Additionally, when $\gamma \cdot \varepsilon$ and η have the same sign, these two sources of bias can cancel each other out. Hence it is possible to have results with approximately zero bias when the front-door conditions do not approximately hold. Finally, we note that for specified values of $\gamma \cdot \varepsilon$ and η, bias-corrected estimates and confidence intervals can be formed in the same manner discussed in VanderWeele and Arah (2011).

3 Application: National JTPA Study

In this section, we compare the performance of the front-door estimator derived in the previous section to the performance of the back-door estimator in the context of the National JTPA Study, a job training evaluation for which we have both experimental data and a nonexperimental comparison group.[6] We measure program impact as the ATT on 18-month earnings in the period post-randomization or post-eligibility. The National JTPA Study is amenable to the use of the front-door estimator because of the presence of nearly one-sided noncompliance.[7]

The National JTPA Study was commissioned by the Department of Labor to gauge the efficacy of the Job Training Partnership Act (JTPA) of 1982. Implemented between November 1987 and September 1989, the National JTPA Study randomized participants at 16 study sites (technically called *service delivery areas*, or SDAs) across the United States into a treatment and control group. Active treatment consisted of being allowed to receive JTPA services following application for the program, while the control group was barred from receiving program services for a period of 18 months following random assignment (Bloom et al., 1993; Orr et al., 1994). The key feature of this study for our analysis is that there was noncompliance among the treated units. In our sample, roughly 57% of adult men and 55% of adult women who were allowed to receive JTPA services

[6] See Appendix B for a more thorough description of the data.

[7] There were a very small number of individuals that received the training program without being assigned to it, however these do not affect the results.

actually utilized JTPA services (see Table 1).

The Study also collected a sample of eligible nonparticipants (ENPs) at 4 service delivery areas as a nonexperimental comparison group. The sample was selected following a screening interview.[8] To match the ENP sample, we restrict the experimental sample to only the 4 sites. Furthermore, we focus on two of five *target groups* defined in the initial study: (1) male adults and (2) female adults.[9] Participants were considered adults if they were at least 22 years old at random assignment. We conduct our analysis separately for the two target groups.

We established the experimental benchmarks by comparing the mean earnings in the 18 months after random assignment of experimental active treatment group sample to the experimental control group sample separately for adult males and adult females. The program impact for adult males was, on average, an increase of $699.60 in the 18-month earnings. For adult females, the impact was $702.17.[10]

Using rich historic data on labor market participation for both the experimental control group and the nonexperimental control group, Heckman et al. (1998) were able to characterize selection bias and thus apply their semiparametric sample selection estimator. As the authors explain, "detailed information on recent labor force status histories and recent earning are essential in constructing comparison groups that have outcomes close to those of an experimental control group" (1020). However, what if such rich data is not available or it is too costly? Are we then unable to create a comparison group that resembles an experimental control group?

Our results from the front-door estimator suggest that even with extremely limited covariates we have recourse to the treated noncompliers in the creation of a comparison group. This was confirmed in Heckman et al. (1997) which shows that no-shows in the National JTPA Study are similar to the experimental control group by calculating their respective conditional probabilities

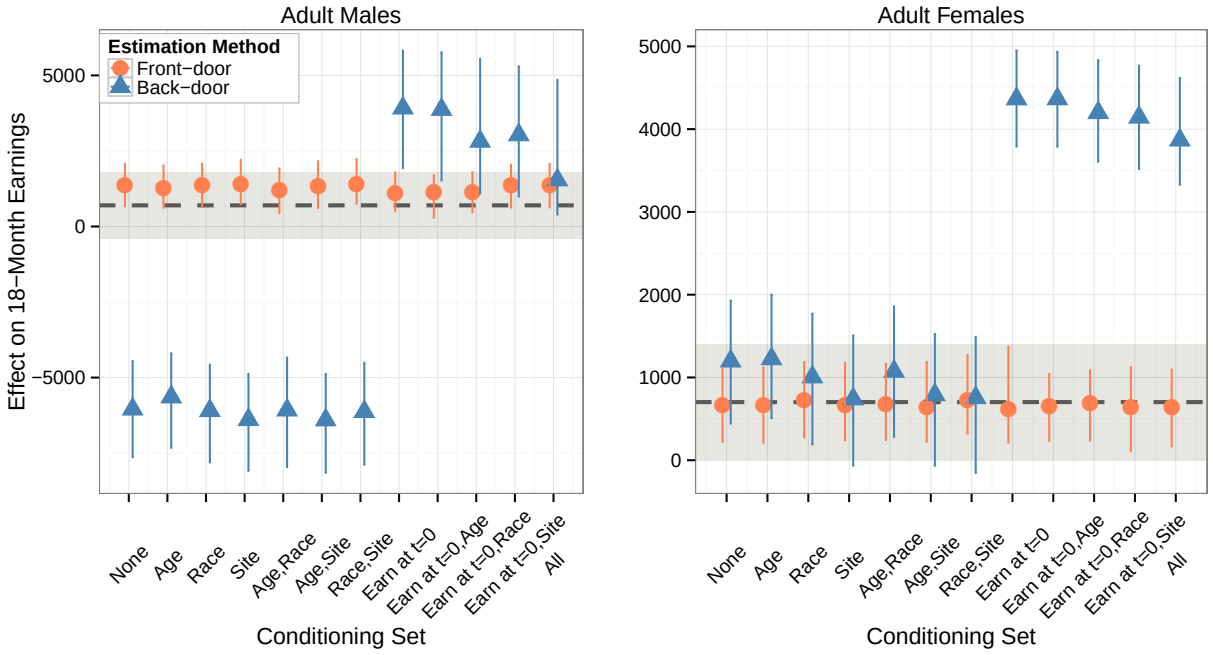
[8] See Appendix B for additional information regarding the ENP sample. See Smith (1994) for details of ENP screening process.

[9] The other 3 target groups were female youths, non-arrestee male youths, and male youth arrestees.

[10] See discussion of how we created our sample and the earnings data in Appendix B.

of being enrollees. We expand on this result here. Figure 1 presents the comparative performance of the front-door and back-door estimators across a variety of simple conditioning sets for adult males and adult females, respectively. We do not assume linearity or additivity in the conditional expectation function $E[Y|a, m, x]$ and thus use kernel-based regularized least squares (KRLS) to obtain three conditional expectations: $E[Y|a_0, M = 0, x]$, $E[Y|a_1, M = 0, x]$, and $E[Y|a_1, M = 1, x]$ (Hainmueller and Hazlett, 2013).[11]

Figure 1: Comparison of Back-door and Front-door Adjustment on JTPA Dataset by Target Group using KRLS. The conditioning sets include permutations of the following variables: age; race dummies for white, black, and other; site dummies; and total earnings in month of random assignment/eligibility screening (RA/ES). The experimental estimate is denoted as a dotted dark grey line, with the shaded grey region representing the 95% confidence interval. 95% bootstrapped percentile confidence intervals for both adjustment methods and the experimental benchmark are based on 10,000 replicates.



The result is striking in that for adult males, the front-door estimates exhibit uniformly less estimation error than the back-door estimates across all the specifications we examine. The error using the null conditioning set from the back-door estimate is -6745.98. This negative error

[11]We report results from KRLS here due to our reluctance to make strong parametric assumptions, but we obtain similar results when using other methods, such as OLS, for estimation.

in the back-door estimate persists even when we condition on age, race, or site. The error in the back-door estimates becomes positive whenever conditioning on the total monthly earnings in the month of random assignment / eligibility screening.[12] The stable performance of the front-door estimates is noteworthy. Without recourse to more detailed data on labor force participation and historic earnings, we find that front-door estimates are preferable to back-door adjustment. The front-door estimates for adult females are similarly stable across specifications. While this result is a bit less striking than for adult males, we would still prefer the front-door estimates compared to the back-door estimates in all but one specification if considering the point estimates (and even in that specification, the back-door estimate has less error only by $1 relative to the front-door estimate).

In sum, we find that for all but a couple of covariate sets, the front-door adjustment has less error than typical back-door adjustment. Moreover, the improvement due to the front-door adjustment is often dramatic, and there is no covariate set where the front-door adjustment has large error. In fact the strong performance of the front-door adjustment relative to the back-door adjustment meant that we were unable to find a hybrid estimator that improved on the front-door approach for this application. Rephrasing our result, we find that using treated units that did not comply and receive JTPA services as proxies for experimental control units yields better estimates than using the nonexperimental control group as the counterfactual for what would have happened to the treated units had they not received treatment. To emphasize this point, we note that it was not actually necessary to collect information on any control units (experimental or nonexperimental) in order to get front-door estimates that are quite close to the experimental benchmark.

3.1 Comparative Sensitivity Analysis for the JTPA

In most applications, we will not have the experimental benchmark presented above. However, using the simplified comparative sensitivity analysis discussed in Section 2.1.1, we show that we would

[12] $t = 0$ is the month of random assignment for the experimental samples and the month of eligibility screening for the nonexperimental control sample.

likely prefer the front-door estimates to the back-door estimates for adult males in this application, even if we did not know the experimental benchmark.

It is helpful to consider how we would react to a simple sensitivity analysis on the back-door estimates for adult males. Suppose we did not have the experimental benchmark or the front-door estimates; we only had available the back-door estimates in Figure 1 (a). Suppose further that we only consider the conditioning sets that include baseline earnings, as these are seen as more credible. If we are willing to assume that a back-door approach would be approximately unbiased if we could measure earning potential as a binary variable U, and if we assume that effects do not depend greatly on the values of the measured covariates, then we can use the simple sensitivity parameters from VanderWeele and Arah (2011). If we think of $U = 1$ individuals as having high earning potential and $U = 0$ individuals as having low earning potential, then γ is clearly positive across all conditioning sets. Additionally, due to the pre-program earnings dip– those that select into treatment have temporarily low baseline earnings at the start of the program– it is likely that $\delta > 0$ and hence that the bias $\gamma \cdot \delta$ is positive.

Now suppose we have the front-door estimates for these sets. We immediately notice that the front-door estimates are all smaller than the back-door estimates. Given that we assume the back-door estimator to have positive bias, the key question is whether the front-door estimator could have enough negative bias so that we would still prefer the back-door estimator. An examination of the front-door sensitivity parameters makes this seem unlikely.

It seems reasonable to assume that the treated non-compliers (signed up, but didn't show up) are generally a bit more diligent and likely to have higher earning potential than the controls (didn't even sign up). This implies that $\varepsilon < \delta$ and $\gamma \cdot \varepsilon < \gamma \cdot \delta$. Furthermore, it is unlikely that there is a direct effect of treatment on the outcome (signing up without showing up should have little effect). This direct effect is the main component of η, although as we discussed previously, we must worry about confounders of the mediator outcome relationship that are not included in the variable U. However, even if we believe that $\eta > 0$ due to such confounders, since we believe the back-door

bias is positive and the front-door estimates are smaller than the back-door estimates, we will prefer the front-door approach for this application as long as $\gamma \cdot \varepsilon > \eta$. This is likely to hold because the "effect" of U on the outcome will dominate the direct "effect" of A for this application, and we expect the front-door imbalance to be non-trivial. Even if $\gamma \cdot \varepsilon < \eta$ we may prefer the front-door approach on absolute bias grounds as long as $|\gamma \cdot \varepsilon - \eta| < \gamma \cdot \delta$.

As with all sensitivity analyses, this analysis is speculative. However, it seems clear from this discussion that the front-door adjustment would possibly have been preferred to the back-door adjustment for this analysis. At the very least, front-door estimates should be presented along with back-door estimates when the conditions discussed above are reasonable.

4 Conclusion

In this paper, we have provided formulas for the large sample bias of front-door adjustments for both Average Treatment Effects on the Treated (ATT) and Average Treatment Effects (ATE). These formulas only utilize potential outcomes in terms of the treatment, and they provide a means for sensitivity analysis with the front-door adjustment. We have further demonstrated that these bias formulas can be compared directly to the bias formulas of VanderWeele and Arah (2011) for standard back-door covariate adjustments. This allows the consideration of when the front-door approach will be preferred to the back-door approach.

In order to provide intuition, we have also presented these bias comparisons in two special cases: the estimation of ATT for nonrandomized program evaluations with one-sided noncompliance and the estimation of ATE using linear structural equation models (in the supplementary material). These comparisons demonstrated that there are broad classes of applications for which the front-door or hybrid adjustments will be preferred to the back-door adjustments. In particular, we illustrated the case of nonrandomized program evaluations with one-sided noncompliance with an application to the National JTPA (Job Training Partnership Act) Study. We show that the front-

door adjustment performs remarkably better than the back-door adjustment over a wide variety of sets of covariates. We also develop a comparative sensitivity analysis that demonstrates the front-door approach likely should have been preferred to the back-door approach even prior to seeing the experimental benchmark.

The results in this paper have implications for research design and analysis. First, the JTPA example demonstrates the importance of collecting post-treatment variables that represent compliance with or uptake of the treatment. This is true even for the analysis of total effects. In this application, the enrollment information was more useful than all other pre-treatment covariates we examined. If such compliance information can be collected, the front-door adjustment should be considered as at least a robustness check for results derived by back-door adjustments. Furthermore, if we have prior beliefs that front-door bias will be smaller than back-door bias, then it may be unnecessary to collect any information on control units. This could be extremely helpful in cases where it is costly to collect pretreatment covariates, or to follow up with the control units to measure outcomes. Finally, we note that this approach provides a method for analysis when it is unethical to withhold treatment from individuals in a study.

A ATT Proofs

A.1 Large-Sample Bias

The bias in the front-door estimate of $E[Y(a_0)|a_1]$ is the following:

$$
\begin{aligned}
B^{fd}_{a_1} &= \mu^{fd}_{0|a_1} - \mu_{0|a_1} \\
&= \sum_x \sum_m P(m|a_0,x) \cdot E[Y|a_1,m,x] \cdot P(x|a_1) \\
&- \sum_x \sum_u E[Y|a_0,x,u] \cdot P(u|x,a_1) \cdot P(x|a_1) \\
&= \sum_x \sum_m P(m|a_0,x) \sum_u E[Y|a_1,m,x,u] \cdot P(u|a_1,m,x) \cdot P(x|a_1) \\
&- \sum_x \sum_u \sum_m E[Y|a_0,m,x,u] \cdot P(m|a_0,x,u) \cdot P(u|a_1,x) \cdot P(x|a_1) \\
&= \sum_x P(x|a_1) \sum_m \sum_u P(m|a_0,x) \cdot E[Y|a_1,m,x,u] \cdot P(u|a_1,m,x) \\
&- \sum_x P(x|a_1) \sum_m \sum_u P(m|a_0,x,u) \cdot E[Y|a_0,m,x,u] \cdot P(u|x,a_1)
\end{aligned}
\tag{15}
$$

The bias of the back-door estimator can be written as the following:

$$
\begin{aligned}
B^{bd}_{a_1} &= \mu^{bd}_{0|a_1} - \mu_0 \\
&= \sum_x E[Y|a_0,x] \cdot P(x|a_1) - \sum_x \sum_u E[Y|a_0,x,u] \cdot P(u|x,a_1) \cdot P(x|a_1) \\
&= \sum_x \sum_u E[Y|a_0,x,u] \cdot P(u|a_0,x) \cdot P(x|a_1) \\
&- \sum_x \sum_u E[Y|a_0,x,u] \cdot P(u|a_1,x) \cdot P(x|a_1) \\
&= \sum_x P(x|a_1) \sum_u E[Y|a_0,x,u] \cdot [P(u|a_0,x) - P(u|a_1,x)]
\end{aligned}
\tag{16}
$$

A.2 Nonrandomized program evaluation with one-sided noncompliance

In the special case of nonrandomized program evaluations with one-sided noncompliance, the front-door and the back-door ATT bias can be written as the following, utilizing the fact that $P(M = 0|a_0, u) = 1$ and $P(M = 0|a_1, u) = 0$ for all u:

$$
\begin{aligned}
B^{fd}_{ATT} &= \mu_1 - \mu^{fd}_{0|a_1} - (\mu_1 - \mu_{0|a_1}) \\
&= \mu_{0|a_1} - \mu^{fd}_{0|a_1} \\
&= -B^{fd}_{a_1} \\
&= \sum_x P(x|a_1) \sum_u E[Y|a_0, M = 0, x, u] P(u|a_1, x) \\
&- \sum_x P(x|a_1) \sum_u E[Y|a_1, M = 0, x, u] P(u|a_1, M = 0, x)
\end{aligned}
$$

Adding and subtracting $\sum_x P(x) \sum_u E[Y|a_0, M = 0, u] \cdot P(u|a_1, M = 0)$:

$$
\begin{aligned}
&= \sum_x P(x|a_1) \sum_u E[Y|a_0, M = 0, x, u] \cdot [P(u|a_1, x) - P(u|a_1, M = 0, x)] \\
&- \sum_x P(x|a_1) \sum_u \{E[Y|a_1, M = 0, x, u] - E[Y|a_0, M = 0, x, u]\} \cdot P(u|a_1, M = 0, x)
\end{aligned}
\tag{17}
$$

A.3 Bias Simplification

In order to improve interpretability of the bias formulas and establish comparability with the results for back-door bias in VanderWeele and Arah (2011), we offer a simplification of the front-door bias formula under one-sided noncompliance. Assuming that U is binary and that quantities do not vary across levels of X, we can rewrite the first term in the final B^{fd}_{ATT} expression above as:

$$
\begin{aligned}
&\sum_x P(x|a_1) \sum_u \mathrm{E}[Y|a_0, M=0, x, u] \cdot [P(u|a_1, x) - P(u|a_1, M=0, x)] \\
&= \sum_x P(x|a_1) E[Y|a_0, M=0, x, U=1] \cdot [P(U=1|a_1, x) - P(U=1|a_1, M=0, x)] \\
&+ \sum_x P(x|a_1) E[Y|a_0, M=0, x, U=0] \cdot [P(U=0|a_1, x) - P(U=0|a_1, M=0, x)] \\
&= \sum_x P(x|a_1) E[Y|a_0, M=0, x, U=1] \cdot [P(U=1|a_1, x) - P(U=1|a_1, M=0, x)] \\
&+ \sum_x P(x|a_1) E[Y|a_0, M=0, x, U=0] \cdot [1 - P(U=1|a_1, x) - (1 - P(U=1|a_1, M=0, x))] \\
&= \sum_x P(x|a_1) E[Y|a_0, M=0, x, U=1] \cdot [P(U=1|a_1, x) - P(U=1|a_1, M=0, x)] \\
&+ \sum_x P(x|a_1) E[Y|a_0, M=0, x, U=0] \cdot [P(U=1|a_1, M=0, x) - P(U=1|a_1, x)] \\
&= \sum_x P(x|a_1) \cdot \{E[Y|a_0, M=0, x, U=1] - E[Y|a_0, M=0, U=0, x]\} \\
&\qquad \cdot [P(U=1|a_1, x) - P(U=1|a_1, M=0, x)]
\end{aligned}
\tag{18}
$$

We thus simplify the front-door bias under one-sided noncompliance to:

$$
\begin{aligned}
B_{ATT}^{fd} = &\sum_x P(x|a_1) \cdot \{E[Y|a_0, M=0, x, U=1] - E[Y|a_0, M=0, U=0, x]\} \\
&\qquad \cdot [P(U=1|a_1, x) - P(U=1|a_1, M=0, x)] \\
&- \sum_x P(x|a_1) \sum_u \{E[Y|a_1, M=0, x, u] - E[Y|a_0, M=0, x, u]\} \\
&\qquad \cdot P(u|a_1, M=0, x)
\end{aligned}
\tag{19}
$$

B National JTPA Study

Our analysis makes use of the following samples in the National JTPA Study: experimental active treatment group, experimental control group, and the nonexperimental / eligible nonparticipant (ENP) group. We restrict our attention to the 4 *service delivery areas* at which the ENP sample

was collected: Fort Wayne, IN; Corpus Christi, TX; Jackson, MS, and Providence, RI. We also only examine 2 target groups: adult males and adult females. Note that the active treatment group for our purposes means receving any JTPA service, even though the services actually received from the JTPA varied across individuals.[13]

The raw data and edited analysis files are available as part of the National JTPA Study Public Use Data from the Upjohn Institute. The covariates for the experimental sample are available through the background information form (BIF) and the covariates for ENPs are available through the long baseline survey (LBS). The experimental samples completed the BIF, which contains demographic information, social program participation, and training and education histories, at the time of random assignment. The ENPs completed the LBS anywhere from 0 to 24 months following eligiblity screening. Unlike the BIF which mostly covers the previous year in terms of labor market experiences, the LBS covers the past 5 years prior to the survey date and thus provides a much richer portrait of labor market participation. Moreover, experimental control units at the 4 ENP sites also received the long baseline survey, completed 1-2 months after random assignment. Heckman et al. (1998), Heckman and Smith (1999), and related works rely on the detailed labor force participation data and earnings histories in LBS to identify selection bias by comparing the experimental control units to the nonexperimental control units. Unfortunately, treated units were never administered the LBS and we have no detailed labor force participation data for multiple years prior to random assignment. Moreover, no one survey instrument was administered to all three of the samples we are using in this analysis, yielding issues of noncomparability. The limited set of covariates we use in the conditioning sets in our analysis have all been established to be comparable by verifying their values across the BIF and LBS for the experimental control group, which completed both surveys at the 4 ENP sites.

The dataset we end up using in our analysis was obtained in communication with Jeffrey Smith

[13]The National JTPA Study classified services received into the following 6 categories: classroom training in occupational skills, on-the-job training, job search assistance, basic education, work experience, and miscellaneous.

and Petra Todd. It is the dataset used in the estimates presented in Section 11 of Heckman et al. (1998) and contains all three samples we use in our analysis. It also contains compliance information for the experimental treated group sample. The covariates we utilize in our analysis have been cross-checked against the raw data from the Upjohn Institute. There are also additional covariates in the Heckman et al. (1998) data that have been imputed using a linear regression as described in Appendix B3 of their paper.

The outcome variable we use in the analysis is total 18-month earnings in the period following random assignment (for experimental units) or eligiblity screening (for ENPs). The monthly total earnings variable available from the public use data files is the `totearn` variable. The data covers months 1-30 after random assignment (denoted as $t+1$ to $t+30$, where t is the time of random assignment). The data also includes data for t, the month of random assignment. Note that this variable is not raw earnings data, but was constructed by Abt Associates from the First and Second Follow-up Surveys, as well as based on data from state unemployment agencies, for the initial JTPA report.[14] Please consult Appendix A of Orr et al. (1994) for description of the First Follow-up Survey, Second Follow-up Survey, and earnings data from state unemployment insurance agencies and Appendix B of the same report for construction and imputation of the 30-month earnings variables. The Narrative Description of the National JTPA Study Public Use Files also contains description of the imputation process (see `http://www.upjohninst.org/erdc/njtpa.html`).

In our analysis, we rely upon the monthly total earnings variable in the dataset we obtained from Jeffrey Smith and Petra Todd. We have verified the earnings data used in the calculation of the program impact from this dataset against the earnings variables in the public use data and they match exactly except for a few individuals where Heckman et al. (1998) have imputed missing monthly data. This applies to around 1% of observations and thus is unlikely to substantively change any results. A unit-by-unit comparison of earnings across the raw data and the data we are using can be

[14] One of the major imputations was a decision to divide raw earnings by a `shares` variable which adjust earnings reported for incomplete months (due to the timing of the interviews) to full monthly earnings.

Table 1: Sample Sizes Before and After Imposing Sample Restrictions. The treated units are broken up into compliers (C) and noncompliers (NC). Control denotes experimental control and ENP denotes the eligible nonparticipants.

	Adult Males				Adult Females			
	Treated		Control	ENP	Treated		Control	ENP
	C	*NC*			*C*	*NC*		
Pre-restriction	843	635	649	667	953	781	830	1340
Post-restriction	834	622	523	384	934	765	706	852

obtained from us upon request.

The full dataset we obtained contains 1478 treated units, 649 experimental control units, and 667 ENPs for adult males. For adult females, there are 1734 treated units, 830 experimental control units, and 1340 ENPs. These numbers already exclude individuals without any earnings records. We follow the sample restrictions in Heckman et al. (1998) to reduce the full dataset to the final sample (see Appendix B1). We impose an age restriction of 22 to 54 years old on the experimental samples to match the ages of the ENP sample. We then omit individuals who are missing data on race and date of eligibility. Finally, we impose a *rectangular restriction* based on quarterly earnings. For experimental control and the ENP samples, we require (i) at least one month of valid earnings prior to random assignment (for experimental controls) or prior to eligibility screening (for ENPs), denoted as $t = 0$, (ii) valid earnings data at $t = 0$, and (iii) at least one month of valid earnings data in months $t + 13$ to $t + 18$. For the treatment group, we impose only restriction iii. The final sample sizes are presented in Table 1.

Even after imposing the rectangular restriction on earnings, some individuals had missing earnings data for some months. In the construction of the 18-month total earnings variable, we mean impute the missing months using the average of the individual's available monthly earnings. Details on the extent of missingness are available from authors upon request.

References

Balke, A. (1995). *Probabilistic counterfactuals: semantics, computation, and applications*. PhD thesis, University of California, Los Angeles. 2

Bloom, H. S., Orr, L. L., Cave, G., Bell, S., and Doolittle, F. (1993). The National JTPA Study: Title IIA Impacts on Earnings and Employment at 18 Months. Bethesda, MD. 10

Chalak, K. and White, H. (2011). Viewpoint: An extended class of instrumental variables for the estimation of causal effects. *Canadian Journal of Economics*, pages 1–51. 2

Cox, D. R. (1960). Regression analysis when there is prior information about supplementary variables. *Journal of the Royal Statistical Society, Ser. B*, 22:172–176. 2

Cox, D. R. and Wermuth, N. (1995). Discussion of 'Causal diagrams for empirical research'. *Biometrika*, 82:688–689. 2

Glynn, A. and Quinn, K. (2011). Why Process Matters for Causal Inference. *Political Analysis*, 19(3):273–286. 2

Hainmueller, J. and Hazlett, C. (2013). Kernel Regularized Least Squares: Moving Beyond Linearity and Additivity Without Sacrificing Interpretability. *MIT Political Science Department Research Paper No. 2012-8*. 12

Heckman, J., Ichimura, H., Smith, J., and Todd, P. (1998). Characterizing selection bias using experimental data. *Econometrica*, 66:1017–1098. 5, 7, 11, 20, 21, 22

Heckman, J., Ichimura, H., and Todd, P. (1997). Matching as an econometric evaluation estimator evidence from evaluating a job training program. *Review of Economic Studies*, 64:605–654. 7, 11

Heckman, J. J., LaLonde, R. J., and Smith, J. A. (1999). The Economics and Econometrics of Active

Labor Market Programs. In Ashenfelter, O. and Card, D., editors, *Handbook of Labor Economics, Volume III*. North Holland. 7

Heckman, J. J. and Smith, J. A. (1999). The pre-programme earnings dip and the determinants of participation in a social programme: implications for simple programme evaluation strategies. *Economic Journal*. 7, 20

Imbens, G. and Rubin, D. (1995). Discussion of 'Causal diagrams for empirical research'. *Biometrika*, 82:694–695. 2

Joffe, M. (2001). Using information on realized effects to determine prospective causal effects. *Journal of the Royal Statistical Society. Series B, Statistical Methodology*, pages 759–774. 2

Kaufman, S., Kaufman, J. S., and MacLehose, R. F. (2009). Analytic bounds on causal risk differences in directed acyclic graphs with three observed binary variables. *Journal of Statistical Planning and Inference*, 139:3473–87. 2

Kuroki, M. and Miyakawa, M. (1999). Identifiability Criteria for Causal Effects of Joint Interventions. *J. Japan Statist. Soc.*, 29(2):105–117. 2

Lee, D. S. (2009). Training, wages, and sample selection: Estimating sharp bounds on treatment effects. *The Review of Economic Studies*, 76(3):1071–1102. 7

Orr, L. L., Bloom, H. S., Bell, S. H., Lin, W., Cave, G., and Doolittle, F. (1994). The National JTPA Study: Impacts, Benefits, And Costs of Title IIA. Bethesda, MD. 10, 21

Pearl, J. (1995). Causal diagrams for empirical research. *Biometrika*, 82:669–710. 2, 3, 4, 5

Pearl, J. (2000). *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 1 edition. 3

Pearl, J. (2009). *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2 edition. 3

Ramsahai, R. (2012). Supplementary variables for causal estimation. In Berzuini, C., Dawid, A., and Bernardinelli, L., editors, *Causal Inference: Statistical Perspectives and Applications*. Wiley and Sons. 2

Shpitser, I. and Pearl, J. (2006). Identification of conditional interventional distributions. Proceedings of the Twenty Second Conference on Uncertainty in Artificial Intelligence (UAI). 2

Smith, J. A. (1994). Sampling Frame for the Eligible Non-Participant Sample. *Mimeo*. 11

Tian, J. and Pearl, J. (2002a). A general identification condition for causal effects. In *Proceedings of the National Conference on Artificial Intelligence*, pages 567–573. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999. 2

Tian, J. and Pearl, J. (2002b). On the identification of causal effects. In *Proceedings of the American Association of Artificial Intelligence*. 2

VanderWeele, T. (2008). The sign of the bias of unmeasured confounding. *Biometrics*, 64(3):702–706. 2

VanderWeele, T. and Robins, J. (2009). Signed directed acyclic graphs for causal inference. *JR Stat Soc B*. 2

VanderWeele, T. J. (2009). On the relative nature of overadjustment and unnecessary adjustment. *Epidemiology*, 20(4):496–499. 2

VanderWeele, T. J. and Arah, O. A. (2011). Bias formulas for sensitivity analysis of unmeasured confounding for general outcomes, treatments, and confounders. *Epidemiology*, 22(1):42–52. 2, 5, 9, 10, 14, 15, 18

Winship, C. and Harding, D. (2008). A Mechanism-Based Approach to the Identification of Age-Period-Cohort Models. *Sociological Methods & Research*, 36(3):362. 2

Zhang, J. L., Rubin, D. B., and Mealli, F. (2009). Likelihood-based analysis of causal effects of job-training programs using principal stratification. *Journal of the American Statistical Association*, 104(485):166–176. 7