Telescope Matching: Reducing Model Dependence in the Estimation of Direct Effects [*]

Matthew Blackwell[†]
Anton Strezhnev[‡]

April 12, 2019

Abstract

Matching methods are widely used to reduce the dependence of causal inferences on modeling assumptions, but their application has been mostly limited to the overall effect of a single treatment. Direct effect analyses, which estimate the effect of a treatment not due to some mediator, have become increasingly popular in the social sciences for a wide variety of inferential targets, including understanding the causal mechanisms of a treatment. Standard matching analyses, however, are not directly applicable to direct effect analyses because of their tendency to induce post-treatment bias, and so almost all applications are dependent on the correct specification of several models. In this paper, we propose a novel two-step matching approach to estimating direct effects, *telescope matching*, that reduces model dependence without inducing post-treatment bias. This method uses matching with replacement to impute missing counterfactual outcomes in a flexible manner and relies on regression models to correct for bias induced by imperfect matches. We show in simulations that our approach is more robust to misspecification of these regression models than non-matching estimators. We derive the asymptotic properties of this estimator and provide a consistent estimator for its variance. Finally, we apply this approach to estimating the direct effect of a job training program on long-term mental health not due to employment and show that it can generate substantively different inferences than standard approaches.

[*]Thanks to Alberto Abadie, Paul Kellstedt, Gary King, Jamie Robins, Jann Spiess, and Yiqing Xu for valuable feedback and discussions. Any remaining errors are our own. Software to implement the methods in this paper will be found in the `DirectEffects` R package on CRAN.

[†]Department of Government and Institute for Quantitative Social Science, Harvard University, 1737 Cambridge St, MA 02138. web: http://www.mattblackwell.org email: mblackwell@gov.harvard.edu

[‡]University of Pennsylvania Law School, 3501 Sansom Street, Philadelphia, PA 19104, web: https://www.antonstrezhnev.com/ email: astrezhn@law.upenn.edu

1 Introduction

Matching is a popular strategy for estimating average treatment effects for a single binary treatment (Rosenbaum, 1995; Dehejia and Wahba, 1999). Its popularity derives from its nonparametric and intuitive nature, but it has been shown to be asymptotically biased due to the possibility of non-exact matches (Abadie and Imbens, 2006). Thus, most applications of matching use a subsequent regression model to adjust for remaining biases (Abadie and Imbens, 2011), and so matching can be viewed as a way to reduce the dependence of inferences on the modeling choices in those regressions (Ho et al., 2006). Because of this appeal, matching have been widely adopted in the social and biomedical sciences.

One setting where matching has made relatively little impact is the estimation of the direct effect of a treatment not due to a mediating variable. These direct effects have become increasingly popular in the social sciences, because they can help to adjudicate theories for why a causal effect of treatment exists (Robins and Greenland, 1992; Imai, Keele, and Yamamoto, 2010; VanderWeele, 2015; Acharya, Blackwell, and Sen, 2016). These quantities of interest may also detect heterogeneous effects that can help to design more effective treatments in the future and help to estimate impulse response functions in time-series and time-series cross-sectional settings (Blackwell and Glynn, 2018).

When there are post-treatment confounders for the mediator-outcome relationship, standard approaches to estimating treatment effects, including matching, are likely to induce post-treatment bias when applied to the estimation of direct effects (Robins, 1997). Thus, several parametric and semi-parametric methods have been developed for estimating these quantities, including the parametric g-formula, structural nested models, and marginal structural models (Robins, 1986; Richardson and Rotnitzky, 2014). Unfortunately, these extant approaches to estimating direct effects require the (correct) specification of several models, meaning our inferences will be heavily dependent on those modeling choices.

In this paper, we present a new matching method for estimating direct effects that helps reduce dependence on these modeling assumptions without inducing post-treatment bias. To do so, we match into steps, first for the mediator and then for the treatment using different covariate sets for

each step. This two-stage approach, which we call "telescope matching," adjusts for both pre- and post-treatment confounders in the first stage, "telescoping out" to only the pre-treatment confounders in the second stage. These matches steps help us impute missing counterfactual outcomes for each unit, which then can be used to estimate direct effects.

We derive the large-sample properties of this matching approach under a fixed number of matches and show that while it is consistent for the controlled direct effect, it possesses a bias due to inexact matches that prevents convergence to a stable asymptotic distribution, as is the case with single-shot matching (Abadie and Imbens, 2006). We thus develop a bias-correction method that uses regression estimators in a similar manner to Abadie and Imbens (2011) and show that, under some regularity conditions, the asymptotic distribution of the bias-corrected and simple matching estimator are the same. We further leverage this bias correction to derive a consistent variance estimator for our estimator. Though we present the regressions here as bias correction for simple matching, telescope matching can also be seen as a way to make regression approaches to direct effects (such as structural nested mean models) more robust to modeling assumptions. We show that this is the case in our simulations—telescope matching has similar performance to these methods when the regression models are correctly specified and shows considerably lower bias when the models are misspecified. Telescope matching has additional benefits in this setting. First, both matching steps can be done and evaluated without access to the outcome, reducing the potential for biased model selection. Second, the matching procedure can be applied to any type of outcome variable, whereas methods like structural nested mean models are difficult to apply to binary outcomes.

This paper proceeds as follows. First, we describe the relevant quantities of interest, including the controlled direct effect, and the assumptions necessary to identify these effects. We then define our telescope matching approach to estimating these direct effects, discuss its large-sample properties, describe the bias-correction approach, and derive variance estimators. Next, we conduct a simulation study that shows how these various estimators perform when a researcher has a correct and incorrect specification of the outcome regression model. Finally, we demonstrate the method in an empirical setting of an experimental study of the effect of job training on mental well-being as mediated by

employment (Huber, 2014).

2 Proposed method

2.1 Notation and assumptions

Let $A_i \in \{0,1\}$ denote values of treatment for unit i. Let $M_i \in \{0,1\}$ denote the value of a post-treatment variable that we seek to hold constant. This may be the same treatment administered a future date or it might be some consequence of the treatment that the analyst believes is part of a causal mechanism. For brevity, we call this variable the mediator since it is post-treatment and may mediate some or all of the effect of treatment. The goal of the analysis is to estimate the effect of treatment on some outcome, Y_i. We define potential outcomes for this under the various combinations of treatment and mediator, $Y_i(a,m)$ (Rubin, 1974; Robins, 1986). For instance, $Y_i(1,1)$ represents the outcome we would see if unit i had been assigned $A_i = 1$ and $M_i = 1$. We make the usual consistency assumption, $Y_i = Y_i(a,m)$ if $A_i = a$ and $M_i = m$, which states that the observed outcome for unit i is the potential outcome for that unit at its observed level of A_i and M_i. Note that, because M_i can be affected by A_i, it too has potential outcomes, $M_i(a)$, that also follow a consistency assumption, $M_i = A_i M_i(1) + (1 - A_i)M_i(0)$.

We define two sets of relevant covariates: baseline and intermediate. The baseline covariates, X_i, are causally prior to both A_i and M_i. Thus, researchers can adjust for these covariates using typical causal inference techniques such as regression, weighting, or matching. The intermediate covariates, Z_i, can be affected by A_i, but are causally prior to M_i and confound the outcome-mediator relationship. These covariates pose problems for standard models when trying to estimate the effect of the treatment and the mediator at the same time due to the potential for post-treatment bias induced by conditioning on them (Rosenbaum, 1984; Robins, 1997).

Our goal in this paper is to estimate the average controlled direct effect (ACDE) (Robins and

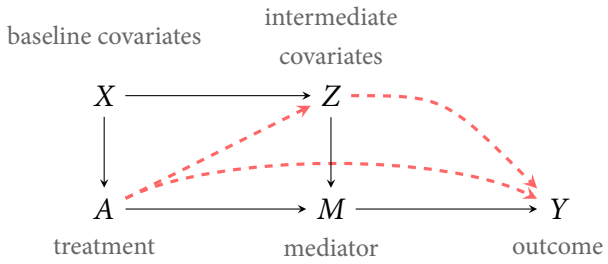

Figure 1: Directed acyclic graph showing the causal relationships in the present setting. Dashed red lines represent the controlled direct effect of the treatment not through the mediator. Unobserved errors are omitted.

Greenland, 1992):

$$\tau = \mathbb{E}[Y_i(1,0) - Y_i(0,0)].$$

This quantity represents the average effect of treatment when the mediator is fixed at a particular value. For this paper, we focus on the ACDE when setting $M_i = 0$, but it is straightforward to extend the discussion and the method to investigate the controlled direct effect at other levels of M_i. We can also define the conditional ACDE:

$$\tau(x) = \mathbb{E}[Y_i(1,0) - Y_i(0,0)|X_i = x]$$

This is the direct effect of treatment within levels of the baseline covariates. We can recover the ACDE from the conditional effects by averaging over the distribution of the data: $\tau = \mathbb{E}[\tau(X_i)]$.

The direct effect stands in contrast to the overall average treatment effect (ATE), which is the difference in average potential outcomes when we just manipulate treatment:

$$\mathbb{E}[Y_i(1) - Y_i(0)] = \mathbb{E}[Y_i(1, M_i(1)) - Y_i(0, M_i(0))].$$

This quantity is the "total" effect of treatment, including both its direct effect and any effects through the mediator. Previous work has focused on the difference between the ATE and the ACDE as evidence for M_i playing a role in the causal mechanism of A_i (Acharya, Blackwell, and Sen, 2016, 2018) and as a measure of how much of the total effect can be eliminated by intervening on M_i (VanderWeele, 2015, pp. 50–52).

We make the following sequential ignorability assumption about the treatment and mediator:

Assumption 1 (Sequential Ignorability). *For every value, a, m, x, z:*

$$\{Y_i(a,m), M_i(a), Z_i(a)\} \perp\!\!\!\perp A_i | X_i = x \tag{1}$$

$$Y_i(a,m) \perp\!\!\!\perp M_i | X_i = x, Z_i = z, A_i = a \tag{2}$$

The first part of this assumption states that the treatment is independent of the potential outcome and the potential values of the mediator, conditional on baseline covariates. The second part states that the mediator is independent of the potential outcomes, conditional on the treatment and the baseline and intermediate covariates. This assumption essentially requires two "selection-on-observables" conditions, one for the treatment and one for the mediator. Thus, there must be no unmeasured confounders for the treatment-outcome relationship after conditioning on X_i and no unmeasured confounders for the mediator-outcome relationship after conditioning on $\{X_i, A_i, Z_i\}$.

We further assume that the distributions of the treatment and mediator are not degenerate at any values of the covariates.

Assumption 2 (Positivity). *For every value, a, x, z, and for some values $\eta > 0$ and $\nu > 0$:*

$$\eta < P(A_i = 1 | X_i = x) < 1 - \eta \tag{3}$$

$$\nu < P(M_i = 1 | X_i = x, Z_i = z, A_i = a) < 1 - \nu \tag{4}$$

The first part of this assumption requires that the treated and control distributions of the baseline covariates have the same support. The second part extends this assumption to the $M_i = 1$ and $M_i = 0$ covariate distributions. These are straightforward generalizations the common support assumptions in the matching literature to the direct effects settings.

A few other pieces of notation will be useful. First, we define a series of conditional expectation functions (CEF) of the potential outcomes, conditional on different sets of covariates. In particular, we define $\mu_{am}(x,z,a) = E[Y(a,m)|X_i = x, Z_i = z, A_i = a]$ and $\mu_{am}(x,a) = E[Y_i(a,m)|X_i = x, A_i = a]$. Let $\mu(x,z,a,m) = E[Y_i|X_i = x, Z_i = z, A_i = a, M_i = m]$ be the CEF of the observed outcome, noting that under Assumption 1, $\mu_{am}(x,z,a) = \mu(x,z,a,m)$ We also define two types of residuals, $\varepsilon_i = Y_i - \mu(X_i, Z_i, A_i, M_i)$ and $\eta_i = \mu_{A_i 0}(X_i, Z_i, A_i) - \mu_{A_i 0}(X_i, A_i)$. The first is the CEF

error for Y_i and the second captures the variation in the CEF of the potential outcomes that is due to Z_i. Given these definitions, we have $E[\eta_i|\mathbf{X}, \mathbf{A}] = 0$ and $E[\varepsilon_i|\mathbf{X}, \mathbf{Z}, \mathbf{A}, \mathbf{M}] = 0$, where $\mathbf{X}$ and $\mathbf{Z}$ are the entire $N \times k_x$ and $N \times k_z$ matrices of pretreatment and intermediate covariates, and $\mathbf{A}$ and $\mathbf{M}$ are the N vectors of the treatment and mediator. Finally, we define various conditional variance functions. Let $\sigma^2(x, z, a, m) = \mathbb{V}[Y_i|X_i = x, Z_i = z, A_i = a, M_i = m]$ and $\sigma^2_\eta(x, a) = \mathbb{V}[E[Y_i(a, 0)|X_i = x, Z_i, A_i = a]|X_i = x, A_i = a] = E[\eta_i^2|X_i = x, A_i = a]$. Again, under Assumption 1, $\sigma^2(x, z, a, m) = \mathbb{V}[Y_i(a, m)|X_i = x, Z_i = z, A_i = a]$.

2.2 The telescope matching procedure

How can we estimate the ACDE? If we simply were to compare the average outcome across levels of the treatment, we will be combining the direct effect and any effect due to changes in the mediator. Instead, we would like to compare the potential outcomes for fixed values of the mediator. To do this, several approaches have been put forward. If both A_i and M_i are randomized, then standard tools for multileveled treatments can be used to estimate the direct effect of treatment since there are no covariates for which to adjust. When there are only baseline confounders, then standard selection-on-observable methods for multi-leveled treatments can be applied (Imbens, 2004). However, when there are post-treatment confounders for the relationship between M_i and Y_i, we must turn to other methods to adjust for this form of confounding.

Our proposed approach, which we call *telescope matching*, imputes values of the missing potential outcomes in a flexible manner. For any particular unit, we only observe one of four possible potential outcomes, an issue sometimes called the fundamental problem of causal inference. To estimate the ACDE when $M_i = 0$, we would like to observe values for $Y_i(1, 0)$ and $Y_i(0, 0)$ for all units. The goal of telescope matching is to use matching methods in order to obtain reasonable imputations of these values for all units.

Let $V_i = (X_i, Z_i)$ be the vector of covariates, both baseline and intermediate. The first step of telescope matching is to match each unit with $M_i = 1$ to some number of units with $M_i = 0$ that have similar values of covariates V_i and identical treatment status A_i. We follow Abadie and Imbens (2006)

in much of our discussion of matching estimators. Given a particular distance metric on the support of V_i (such as the Euclidean norm or the Mahalanobis distance) and given a particular unit with $M_i = 1$, we choose L units, here indexed by ℓ, that are the closest to i in terms of covariate distance that have $M_\ell = 0$. Let $\mathcal{J}_L^m(i)$ denote this set of units that are matched to some unit i with $M_i = 1$. Matching is done with replacement so a control unit might be matched to multiple treated units and we let $K_L^m(i) = \sum_{k=1}^N \mathbb{I}\{i \in \mathcal{J}_L^m(k)\}$ be the number of times that unit i is used as a match in stage, where $\mathbb{I}\{\cdot\}$ is the indicator function. As in Abadie and Imbens (2006), this quantity is important to the asymptotic distribution of the matching estimator.

Typically, matching would be used to estimate the effect of M_i on Y_i, but here we are actually more interested in obtaining a good estimate of the potential outcome under $M_i = 0$ for all units, including those observed to have $M_i = 1$. Let $Y_i(A_i, 0)$ be the potential outcome under the observed treatment status for i, but with the mediator set to 0, which is unobserved for any unit with $M_i = 1$. We define the following imputation:

$$\widehat{Y}_i(A_i, 0) = \begin{cases} Y_i & \text{if } M_i = 0 \\ \frac{1}{L} \sum_{\ell \in \mathcal{J}_L^m(i)} Y_\ell & \text{if } M_i = 1 \end{cases}$$

For units observed with $M_i = 0$, we observe $Y_i(A, 0) = Y_i$ by consistency. However, for units with $M_i = 1$, we need to impute the missing counterfactual outcome and do so by averaging the outcome among those units with $M_i = 0$ which were matched to unit i. These units have identical treatment levels A_i and are the closest to i in terms of the baseline and intermediate covariates.

In the second stage, we match each unit to L units of the opposite treatment status with similar values of the baseline covariates as if we were attempting to estimate the ATE of A_i adjusting for X_i. We could allow for different matching ratios in the two stages, but for simplicity, we focus on the case where the ratios are the same. Let $\mathcal{J}_L^a(i)$ be the indices of the units matching to treated unit i such that $A_j = 1 - A_i$ for all $j \in \mathcal{J}_L^a(i)$ and $K_L^a(i) = \sum_{j=1}^N \mathbb{I}\{i \in \mathcal{J}_L^a(j)\}$ be the number of times i is used as a match in the A_i stage. Note that here, as opposed to in the first stage, we are matching treated to control and control to treated to ensure that we can estimate the overall ACDE (rather than the ACDE conditional on the treated).

Once we have both matching solutions, we can generate imputations of the relevant potential outcomes for the ACDE by treating the first-stage imputations, $\widehat{Y}_i(A_i, 0)$, as the outcomes for the second-stage matching. That is, we use the second-stage to impute for the potential outcome under treatment with $M_i = 0$:

$$\widehat{Y}_i(1,0) = \begin{cases} \widehat{Y}_i(A_i, 0) & \text{if } A_i = 1 \\ \frac{1}{L} \sum_{j \in \mathcal{J}_L^a(i)} \widehat{Y}_j(A_j, 0) & \text{if } A_i = 0 \end{cases}$$

We can also define similar values for the potential outcomes under control:

$$\widehat{Y}_i(0,0) = \begin{cases} \frac{1}{L} \sum_{j \in \mathcal{J}_L^a(i)} \widehat{Y}_j(A_j, 0) & \text{if } A_i = 1 \\ \widehat{Y}_i(A_i, 0) & \text{if } A_i = 0 \end{cases}$$

With these definitions in hand, we can then apply a standard difference in means matching estimator. In particular, the simple telescope matching estimate of the ACDE then becomes,

$$\widehat{\tau} \equiv \frac{1}{N} \sum_{i=1}^{N} \left(\widehat{Y}_i(1,0) - \widehat{Y}_i(0,0) \right).$$

Both $K_L^m(i)$ and $K_L^a(i)$ tell us how much unit i is contributing to the overall estimate through being matched in the first and second stages, respectively. Of course, units with $M_i = 0$ might also contribute *indirectly* if they are matched to a $M_i = 1$ unit in the first stage and that $M_i = 1$ unit is used as a match in the second stage. To account for such indirect contributions of a unit, let $K_L^{am}(i) = \sum_{j=1}^{N} \mathbb{I}\{i \in \mathcal{J}_L^m(j)\} K_L^a(j)$ be the number of times a first-stage match with $M_i = 0$ is implicitly used as a match in the second stage because the unit to which it was matched is selected as a match in the second stage. It is possible to rewrite this simple telescope matching estimator as a weighted average of outcomes for units with $M_i = 0$, with these "number of times matched" values contributing to the weights:

$$\widehat{\tau} = \frac{1}{N} \sum_{i=1}^{N} (2A_i - 1)(1 - M_i) \left(1 + \frac{K_L^a(i)}{L} + \frac{K^m(i)}{L} + \frac{K_L^{am}(i)}{L^2} \right) Y_i \tag{5}$$

This weighted-average version of the estimator highlights how the K terms might affect the variance of our estimators—units that are used as matches many times can lead to large weights and thus higher

variances. This provides one reason to keep the number of matches L relatively low. Below we discuss how this relates to inverse probability of treatment weighting.

2.3 Bias and consistency

To understand the benefits and potential drawbacks of such a matching approach, we investigate the large-sample properties of this estimator. Abadie and Imbens (2006) showed that in the context of estimating the overall ATE, the equivalent simple matching procedure was biased due to imperfect matches. Further, they showed that with a fixed size for the matched set, L, this bias converges to 0 as the sample size increased, but at a rate slow enough to affect the asymptotic normality of the matching estimator. In this section, we show that a similar account holds in the present setting.

In the Supplemental Materials, we show that one can decompose the estimation error of $\widehat{\tau}$ as follows:

$$\widehat{\tau} - \tau = \left(\frac{1}{N} \sum_{i=1}^{N} \tau(X_i) - \tau \right) + E_L^m + E_L^a + B_L^m + B_L^a \tag{6}$$

The first term in the decomposition, $(1/N) \sum_{i=1}^{N} \tau(X_i) - \tau$, is the difference between the sample average of the conditional ACDEs and the true ACDE, which converges to o under a standard law of large numbers. Next in the decomposition are two weighted sums of the residuals:

$$E_L^m = \frac{1}{N} \sum_{i=1}^{N} (2A_i - 1)(1 - M_i) \left(1 + \frac{K_L^a(i)}{L} + \frac{K_L^m(i)}{L} + \frac{K_L^{am}(i)}{L^2} \right) \varepsilon_i$$

$$E_L^a = \frac{1}{N} \sum_{i=1}^{N} (2A_i - 1) \left(1 + \frac{K_L^a(i)}{L} \right) \eta_i$$

The error due to the first-stage is mean-zero conditional on all variables, $\mathbb{E}[E_L^m | \mathbf{X}, \mathbf{A}, \mathbf{Z}, \mathbf{M}] = 0$, and the error due to the second-stage is mean-zero conditional on the baseline covariates and the treatment, $\mathbb{E}[E_L^a | \mathbf{X}, \mathbf{A}] = 0$. Thus, the first three terms impose no bias on the matching estimator.

Finally, the last two terms capture the bias of the matching procedure due to the first and second-

stages of matching:

$$B_L^m = \frac{1}{N}\sum_{i=1}^{N}(2A_i - 1)M_i \left(1 + \frac{K_L^a(i)}{L}\right)\left(\frac{1}{L}\sum_{\ell \in \mathcal{J}_L^m(i)} \mu_{A_i 0}(X_\ell, Z_\ell, A_i) - \mu_{A_i 0}(X_i, Z_i, A_i)\right)$$

$$B_L^a = \frac{1}{N}\sum_{i=1}^{N}(2A_i - 1)\left[\frac{1}{L}\sum_{j \in \mathcal{J}_L^a(i)} \mu_{1-A_i,0}(X_i, 1 - A_i) - \mu_{1-A_i,0}(X_j, 1 - A_i)\right]$$

These bias terms reflect the matching discrepancy at each stage of matching. For instance, the last term in the definition of B_L^m is the difference in the expectation of the outcome for the covariates for unit i and for the units matched to i. This bias is amplified by the number of times that this $M_i = 1$ unit is matched in the second stage. If the matches were perfect, then we would have $X_i = X_\ell$ and $Z_i = Z_\ell$ for all $\ell \in \mathcal{J}_L^m(i)$ and $X_i = X_j$ for all $j \in \mathcal{J}_L^a(i)$, and both of these bias terms would be equal to 0. In general, however, matches are imperfect when we have any continuous covariates and so these bias terms will not be mean-zero (Abadie and Imbens, 2006). Importantly for the results below, though, these values do converge to 0 as N increases.

To establish the large-sample properties of the matching estimator, we make the following regularity conditions, which mostly generalize those of Abadie and Imbens (2006) to the current setting.

Assumption 3 (Regularity conditions). *We assume the following:* (i) *Let* $V_i = (Z_i, X_i)$ *be a random vector of* $k = k_z + k_x$ *continuous covariates distributed on* $\mathbb{R}^k$ *with compact and convex support* $\mathbb{V}$*, with its density bounded and bounded away from zero.* (ii) $\{(Y_i, M_i, Z_i, A_i, X_i)\}_{i=1}^N$ *are independent and identically distributed.* (iii) *The functions* $\mu(x, z, a, m)$*,* $\sigma^2(x, z, a, m)$*, and* $\sigma_\eta^2(x, a)$ *are Lipschtiz on* $\mathbb{V}$*.* (iv) $E[Y_i^4 | V_i = v, A_i = a, M_i = a]$ *exists and is uniformly bounded in* $\mathbb{V}$*.* (v) $\sigma^2(x, z, a, m)$ *and* $\sigma_\eta^2(x, a)$ *are bounded away from o.*

These assumptions conditions impose smoothness on conditional expectations and variances as functions of the covariates and ensure that sufficient moments of the outcome exist to allow for convergence in distribution. These conditions ensure that even though the simple matching estimator is biased, it is consistent for the true ACDE.

Theorem 1. *Suppose that Assumptions 1, 2, and 3 hold. Then, (i)* $\widehat{\tau} - \tau \xrightarrow{p} 0$ *and (ii)* $\sqrt{N}\left(\widehat{\tau} - B_L^m - B_L^a\right) \xrightarrow{d} N(0, \sigma^2)$, *where,* $\sigma^2 = V^{\tau(X)} + V^{\eta} + V^{\varepsilon}$, *and*

$$
\begin{aligned}
V^{\tau(X)} &= \mathbb{E}[(\tau(X_i) - \tau)^2], \\
V^{\eta} &= \mathbb{E}\left[\left(1 + \frac{K_L^a(i)}{L}\right)^2 \sigma_\eta^2(X_i, A_i)\right], \\
V^{\varepsilon} &= \mathbb{E}\left[(1 - M_i)\left(1 + \frac{K_L^a(i)}{L} + \frac{K_L^m(i)}{L} + \frac{K_L^{am}(i)}{L^2}\right)^2 \sigma^2(X_i, Z_i, A_i, 0)\right].
\end{aligned} \tag{7}
$$

Proofs for all results are in the Supplemental Materials. The crux of part (i) of this result comes from the fact that the terms in the decomposition in equation 6 all converge to o in probability. Unfortunately, without further assumptions, the bias terms dominate the distribution of the estimator as $N \to \infty$, so that the simple matching estimator will not converge in distribution at the $\sqrt{N}$ rate (Abadie and Imbens, 2006). The second part of this theorem shows that when the bias terms are removed, the matching estimator is asymptotically normal with a variance that depends on the distribution of the number of times a unit is used as a match. Even though these results ignore the bias terms, they are still useful because the bias correction that we describe next will converge at a fast enough rate so it can be ignored asymptotically (Abadie and Imbens, 2011).

2.4 Bias correction

Due to large-sample bias of a simple matching estimator, Abadie and Imbens (2011) proposed a bias-corrected estimator that estimates and removes the bias from a simple matching procedure. In this section, we extend this idea to the present two-stage setting. In particular, we propose estimating the two bias terms with regression estimators of the two relevant CEFs, $\widehat{\mu}(x, z, a, m)$ and $\widehat{\mu}_{a0}(x, a)$. As in Abadie and Imbens (2011), we define a flexible series estimator that grows more complex with the sample size. Let $\lambda = (\lambda_1, \ldots, \lambda_k)$ be a k-dimensional vector of nonnegative integers with order $|\lambda| = \sum_{i=1}^k \lambda_i$. This vector will define a polynomial in x, so that $x^\lambda = x_1^{\lambda_1} \cdots x_k^{\lambda_k}$. We collect these vectors into a series that is nondecreasing in its order. That is, we let $\{\lambda(G)\}_{G=1}^{\infty}$ be a series of all distinct λ vectors such that $|\lambda(G)|$ is nondecreasing. Let $p_G(x) = x^{\lambda(G)}$ be the polynomial induced by

the Gth vector in this series. Let $p^G(x) = (p_1(x), \ldots, p_G(x))'$ be the vector of all such polynomials up to G. We now define two nonparametric series estimators for the two conditional expectations of interest:

$$\widehat{\mu}(x,z,a,0) = p^{G(N)}(x,z)' \left(\sum_{i:A_i=a,M_i=0} p^{G(N)}(V_i) p^{G(N)}(V_i)' \right)^{-} \left(\sum_{i:A_i=a,M_i=0} p^{G(N)}(V_i) Y_i \right),$$

$$\widehat{\mu}_{a0}(x,a) = p^{G(N)}(x)' \left(\sum_{i:A_i=a} p^{G(N)}(X_i) p^{G(N)}(X_i)' \right)^{-} \left(\sum_{i:A_i=a} p^{G(N)}(X_i) \widetilde{Y}_i(A_i,0) \right),$$

where $\widetilde{Y}_i(A_i,0)$ is a bias-corrected imputation defined below, and $(\cdot)^-$ represents a generalized inverse. Essentially, these are linear regressions of the outcome (or transformed outcome) on a vector of polynomials of the appropriate covariates for that regression. Furthermore, the polynomials grow more complex as the sample size grows, which ensures that this approximation will converge to the true expectation under the appropriate smoothness conditions. The first of these regresses Y_i on X_i and Z_i within levels of A_i and M_i to estimate $\widehat{\mu}(X_i, Z_i, A_i, 0)$. For the second-stage regression $\widehat{\mu}_{A_i 0}(X_i, A_i)$, we use this first-stage regression and the first-stage matching to create a bias-corrected imputation of the missing potential outcome, $Y_i(A_i, 0)$:

$$\widetilde{Y}_i(A_i,0) = \begin{cases} Y_i & \text{if } M_i = 0 \\ \frac{1}{L} \sum_{\ell \in \mathcal{J}_L^m(i)} \left(Y_j + \widehat{\mu}(X_i, Z_i, A_i, 0) - \widehat{\mu}(X_\ell, Z_\ell, A_\ell, 0) \right) & \text{if } M_i = 1 \end{cases}$$

We then treat that imputation as the dependent variable and regress it onto the baseline covariates.

These series estimators will provide a good approximation to the true CEFs when certain conditions are met, as summarized by Assumption 4.

Assumption 4 (Smoothness for bias correction)**.** *We assume the following:* (i) $G(N) = O(N^v)$ *with* $0 < v < \min(2/(4k+3), 2/(4k^2-k))$*;* (ii) *there are constants* $C_1 > 0$ *and* $C_2 > 0$ *such that for each* λ *with* $|\lambda| = k$*, the derivatives* $\partial^\lambda \mu(x,z,a,0)$ *and* $\partial^\lambda \mu_{a0}(x,a)$ *exist and satisfy* $\sup_{(x,z)\in\mathbb{V}} |\partial^\lambda \mu(x,z,a,0)| < C_1$ *and* $\sup_{x\in\mathbb{X}} |\partial^\lambda \mu_{a0}(x,a)| < C_2$.

Assumption 4(i) places bounds on how quickly the complexity of the series estimator can grow, while Assumption 4(ii) ensures that the underlying CEFs are sufficiently smooth to be well-approximated

by the series estimator. In our simulations and empirical examples below, we use a simple linear, additive regression model, which is a type of series estimator. Other flexible estimation methods such as regression trees may also perform well in this context, but these series estimators have the benefit of taking advantage of the assumed smoothness in the CEFs.

We can then use these regressions to obtain estimates of the bias terms themselves:

$$\widehat{B}_L^m = \frac{1}{N}\sum_{i=1}^{N}(2A_i - 1)\left(1 + \frac{K_L^a(i)}{L}\right) M_i \left(\frac{1}{L}\sum_{\ell \in J_L^m(i)} \widehat{\mu}(X_\ell, Z_\ell, A_i, 0) - \widehat{\mu}(X_i, A_i, Z_i, 0)\right)$$

$$\widehat{B}_L^a = \frac{1}{N}\sum_{i=1}^{N}(2A_i - 1)\left[\frac{1}{L}\sum_{j \in J_L^a(i)} \widehat{\mu}_{1-A_i,0}(X_i, 1 - A_i) - \widehat{\mu}_{1-A_i,0}(X_j, 1 - A_i)\right]$$

If the regression estimators are consistent for their respective CEFs, then $\widehat{B}_L^m$ and $\widehat{B}_L^a$ converge in probability to the bias terms B_L^m and B_L^a, respectively. With these estimates in hand, we define the following bias-corrected telescope matching estimator:

$$\widetilde{\tau} = \widehat{\tau} - \widehat{B}_L^m - \widehat{B}_L^a. \tag{8}$$

We establish the asymptotic distribution of this estimator in Theorem 2.

Theorem 2 (Bias-corrected matching). *Suppose that Assumptions 1-4 hold. Then,*

$$\sqrt{N}\left(B_L^m + B_L^a - (\widehat{B}_L^m + \widehat{B}_L^a)\right) \xrightarrow{p} 0, \qquad \textit{and}$$

$$\sqrt{N}(\widetilde{\tau} - \tau) \xrightarrow{d} N(0, \sigma^2).$$

Theorem 2 shows that, under the above smoothness conditions, the bias-correction terms converge at a rate faster than $\sqrt{N}$ and that using the estimated bias rather than the true bias does not affect the large-sample distribution of the matching estimator. Of course, this might understate the sampling variance in finite samples where the variance of the bias correction estimator is non-negligible. In particular, the bias terms may converge more slowly when we are interested in the ACDE with the mediator at $M_i = 0$, but there are many $M_i = 1$ units in the data since the bias due to the first matching step, B_L^m, since, intuitively, there is more bias to correct in this setting. Furthermore, the estimation error in the bias correction, $\widehat{B}_L^m - B_L^m$, will be correlated with the residuals here E_L^m in part due to the

shared value of the weights in these two terms. As Theorem 2 shows, this correlation will converge to zero at a rate faster than $\sqrt{N}$, but in these situations σ^2 might not be a good approximation to the finite sample variance of $\widetilde{\tau}$.

2.5 Inference

Conducting inference for telescope matching requires a valid method for estimating standard errors. Matching with replacement, as we propose here, complicates variance estimation because it creates dependence across the imputed counterfactuals, leading to the complicated form of the variance of the matching estimator in (7). We consider three approaches to variance estimation: (1) an asymptotic variance estimator based on (7), (2) a standard non-parametric boostrap, and (3) the weighted bootstrap as proposed by Otsu and Rai (2017).

One approach to estimating the variance of the matching estimator is to directly estimate the components of (7), which are the variance of the conditional ACDEs and weighted averages of the conditional variances of the outcomes. A straightforward way to implement such an estimator is to replace the populations quantities with their sample counterparts, with estimators for the conditional variances: $\widehat{\sigma}^2 = \widehat{V}^{\tau(X)} + \widehat{V}^{\eta} + \widehat{V}^{\varepsilon}$, where:

$$\begin{aligned}
\widehat{V}^{\tau(X)} &= \frac{1}{N}\sum_{i=1}^{N} \left(\widehat{\mu}_{10}(X_i) - \widehat{\mu}_{00}(X_i) - \widetilde{\tau}\right)^2, \\
\widehat{V}^{\eta} &= \frac{1}{N}\sum_{i=1}^{N} \left(1 + \frac{K_L^a(i)}{L}\right)^2 \left(\widehat{\mu}(X_i, Z_i, A_i, 0) - \widehat{\mu}_{A_i 0}(X_i, A_i)\right)^2, \\
\widehat{V}^{\varepsilon} &= \frac{1}{N}\sum_{i=1}^{N} (1 - M_i) \left(1 + \frac{K_L^a(i)}{L} + \frac{K_L^m(i)}{L} + \frac{K_L^{am}(i)}{L^2}\right)^2 \left(Y_i - \widehat{\mu}(X_i, Z_i, A_i, 0)\right)^2.
\end{aligned} \tag{9}$$

This estimator relies on the estimators for the conditional expectations that we also use for the bias-correction. Alternatively, one could use a matching approach to estimate these conditional variances as in Abadie and Imbens (2006), though these matching estimators can often be improved using bias correction techniques that lead to a similar estimator as presented here. In the next theorem, we show that the same assumptions that justify the bias correction also imply this variance estimator will be consistent for the asymptotic variance of the matching estimator.

Theorem 3 (Variance estimator). *Suppose that Assumptions 1-4 hold. Then,* $\widehat{\sigma}^2 \xrightarrow{p} \sigma^2$.

While the bootstrap is a popular approach for many methods, it is well known that conventional non-parametric bootstrapping, resampling observations $\{Y_i, X_i, Z_i, A_i, M_i\}$, is invalid for matching estimators (Abadie and Imbens, 2008). This is due to the inability of the naive boostrap to preserve the distributions of $K_m^a(i)$, the counts of the number of times unit i is used as a match, across resamples. In the case of telescope matching, the same issue persists for the other match counts: $K_L^m(i)$ and $K_L^{am}(i)$.

Recently, Otsu and Rai (2017) proposed a method for using a variety of bootstrap techniques in the matching setting. They show that when the bias-corrected matching estimator is written in a linearized form such that $\widetilde{\tau} = \sum_{i=1}^N \widetilde{\tau}_i$ where $\widetilde{\tau}_i$ consists only of functions of observation i, one could use a weighted bootstrap of the residuals, $\widetilde{\tau}_i - \widetilde{\tau}$, to obtain valid confidence intervals for matching estimators. This "weighted" boostrap resamples the ith contribution to the overall estimate rather than resampling units and matching again in the resampled units. We show in the Supplemental Materials that it is possible to write the contribution of the ith observation to our bias-corrected estimator as:

$$\begin{aligned}\widetilde{\tau}_i = (2A_i - 1)\Bigg[&(1 - M_i)\left(1 + \frac{K_L^a(i)}{L} + \frac{K_L^m(i)}{L} + \frac{K_L^{am}(i)}{L^2}\right) Y_i \\ &- \left((1 - M_i)\left(\frac{K_L^m(i)}{L} + \frac{K_L^{am}(i)}{L^2}\right) - M_i\left(1 + \frac{K_L^a(i)}{L}\right)\right)\widehat{\mu}_{A_i,0}(X_i, A_i, Z_i) \\ &- \left(\widehat{\mu}_{1-A_i,0}(X_i, 1 - A_i) + \frac{K_L^a(i)}{L}\widehat{\mu}_{A_i,0}(X_i, A_i)\right)\Bigg] \quad (10)\end{aligned}$$

Otsu and Rai (2017) shows that one can bootstrap the sampling distribution, $(\widetilde{\tau} - \tau)$ with

$$T^* = \frac{1}{N}\sum_{i=1}^N W_i^*(\widetilde{\tau}_i - \widetilde{\tau}),$$

where W_i^* are random variables that satisfy a few basic regularity conditions (for more details, see Appendix A of Otsu and Rai, 2017). For our purposes, we use the wild bootstrap (Mammen, 1993), which draws the $W_i^* = \nu_i$, where ν_i is random draw from the two-point distribution that takes on the value $-(\sqrt{5}-1)/2$ with probability $(\sqrt{5}+1)/2\sqrt{5}$ and $(\sqrt{5}+1)/2$ with probability $(\sqrt{5}-1)/2\sqrt{5}$. As discussed by Otsu and Rai (2017), this approach avoids the issues with the naive row-resampling

bootstrap method that is commonly used and analyzed by Abadie and Imbens (2008) in the matching context. In particular, the weighted bootstrap ensures that the distribution of the $K_L^a(i)$, $K_L^m(i)$, and $K_L^{am}(i)$ are preserved across resamples.

Finally, we note that both the weighted bootstrap and our asymptotic variance estimator target σ^2, the variance of the simple matching estimator. This is justified by Theorem 2, which shows that the bias-corrected estimator will have the same asymptotic variance as the simple matching estimator. However, in small samples, the variation in $\widetilde{\tau}$ due to the bias correction estimation will be non-negligible, but will be ignored by both of these approaches. One advantage of the naive bootstrap in this case is that it will account for both the matching and bias-correction estimation uncertainty. To assess how these three methods perform in practice, we evaluate them in a simulation study with varying sample sizes in Section 3.

2.6 Relationship to other approaches

Controlled direct effects have been the focus of a great deal of statistical and empirical studies over the last few decades. As pointed out by Robins (1986) and Rosenbaum (1984), these direct effects are not identified from standard approaches that condition on M_i and Z_i due to the potential post-treatment bias, which is sometimes called collider bias. This has been interpreted as preventing the application of standard regression or matching estimators to this setting. Instead, estimation of these effects has focused on three general approaches: (1) parametric g-estimation, (2) structural nested models, and (3) inverse probability of treatment weighting (IPTW). Parametric g-estimation exploits the fact that the mean of the potential outcomes can be identified by integrating over the distribution of the post-treatment covariates, Z_i. This integration, though, requires parametric models for the outcome and for joint distribution of the covariates, which can be very demanding when there are more than a handful of covariates.

Structural nested mean models (SNMMs) focus on modeling the effects of both the treatment and the mediator, and their implementation requires models for the outcome conditional on the covariates or the treatment and mediator conditional on the covariates (Robins, 2000). Sequential

ignorability implies a particular estimating equation approach that can be used to estimate the controlled direct effect, but these models can also be seen as an imputation approach. In particular, a SNMM captures the effect of the mediator in "blip-down" or "demediation" function:

$$\gamma_m(x, z, a) = E[Y_i(a, 1) - Y_i(a, 0)|X_i = x, A_i = a, Z_i = z, M_i = 1]$$

In the binary context, this function models the (conditional) average treatment effect on the treated of M_i and is parameterized in terms of the covariates and levels of treatment, a, which can permit effect modification. For example, we might have:

$$\gamma_m(x, z, a; \beta) = \beta_0 + \beta_1 a.$$

Here, the effect of the mediator would be β_0 when $A_i = 0$ and $\beta_0 + \beta_1$ when $A_i = 1$. Similar to our matching approach, SNMMs can be seen as using the following imputation for the missing potential outcomes under $M_i = 0$:

$$\overline{Y}_i(A_i, 0) = \begin{cases} Y_i & \text{if } M_i = 0 \\ Y_i - \widehat{\gamma}_m(X_i, Z_i, A_i) & \text{if } M_i = 1 \end{cases}$$

Thus, SNMMs rely on a consistent estimate of the (possibly heterogeneous) causal effects of M_i, which will either rely on a correctly specified model for M_i or an outcome regression model such as the one used in the bias correction above. When linear regression models for the outcomes are used to estimate the parameters of the γ function, this approach has been called *sequential g-estimation*, and is a natural comparison to our bias-corrected estimator which also relies on linear outcome models. One additional disadvantage of SNMMs is that they are difficult to apply to binary outcomes (Robins, 2000; Robins and Rotnitzky, 2004), whereas the matching approach here does not depend on the support of the outcome, though the bias correction may perform better on continuous outcomes.

Inverse probability of treatment weighting (IPTW) represents the third popular approach to estimating direct effects. In the current setting, it would require consistent estimates of the probability of the mediator given the treatment and intermediate and baseline covariates, $e_m(x, z, a) = \Pr(M_i = 1|X_i = x, Z_i = z, A_i = a)$, and the probability of treatment given the baseline covariates,

$e_a(x) = \Pr(A_i = 1 | X_i = x)$. Weights within levels of A_i for each unit can then be constructed as $W_i(a) = \frac{M_i}{e_m(X_i, Z_i, a)} + \frac{1-M_i}{1-e_m(X_i, Z_i, a)}$, so that the weight for unit i is the inverse of the probability of receiving the level of M_i it actually had. If we have consistent estimates of these functions, weighting the data by the above weights will remove the confounding (due to X_i and Z_i) of the relationship between M_i and Y_i. This allows M_i to simply be included as an additional control in any matching, weighting, or regression approach to estimating the (direct) effect of A_i on Y_i. Unfortunately, in practice, IPTW can have poor performance due to unstable weights when the probability of $M_i = 1$ is close to 0 or 1, which can be compounded by model misspecification. As shown in (5), our simple telescope matching estimator can also be thought of as a weighting estimator, where the weights are determined by the number of times a unit is used as a match. The IPTW approach leverages a correctly specified model for the propensity scores, whereas the matching approach attempts to directly construct weights based on minimizing imbalance across levels of A_i and M_i in the covariate distribution of the data.

In addition, a host of *doubly robust* methods have been developed that combine features of the SNMM and weighting approaches. These methods require models for both (a) the outcome-covariate relationship and (b) the mediator/treatment propensity scores. These methods are doubly robust in the sense that they are consistent for direct effects when either (a) or (b) are correctly specified. With telescope matching on the other hand, we model the outcome then use matching to help guard against misspecification.

Finally, a few papers have implemented a similar sequential matching approach to estimate controlled direct effects. Lechner and Miquel (2010) implement such a strategy, but focus on sequentially matching on the propensity scores for A_i and M_i and do not derive formal properties of that estimator. Our approach differs in that we match directly on distances in the covariate space, but it is likely possible to extend our theoretical results to propensity score matching in a similar vein to Abadie and Imbens (2016). More recently, Huber, Lechner, and Strittmatter (2018) use a matching approach to estimate controlled direct effects in a setting where there are no intermediate confounders.

3 Simulation study

We evaluate the performance of telescope matching against existing direct effect methods using a simulation in which we artificially introduce model misspecification. We show that while the performance of telescope matching remains comparable to sequential g-estimation when the true model is known, it is much more robust when the outcome models are incorrectly specified. This simulation follows an approach similar to that of Kang and Schafer (2007) which considered the robustness of "doubly robust" estimators in situations where the functional form of the outcome-confounder relationship was not known.

Our assumed data generating process, which reflects a common situation encountered by researchers, is as follows. We have two observed pre-treatment confounders, X_{i1} and X_{i2}, both $\mathcal{N}(0,1)$, and one post-treatment confounder Z_i. The probability unit i receives treatment is:

$$Pr(A_i = 1|X_{i1}, X_{i2}) = \frac{1}{1 + \exp(-(-X_{i1} + .5X_{i2}))}$$

The post-treatment confounder, Z_i, is a function of treatment and of another, unobserved, confounding factor affecting both Z_i and outcome Y_i. Therefore, while Z_i is causally affected by treatment A_i, it itself does not directly affect Y_i. Rather, it is a control variable that can block confounding due to the unobserved common cause, denoted U_i. Our specific form for Z_i is chosen as:

$$Z_i = .5A_i + \gamma_i + \delta U_i$$

where $\gamma_i \sim \mathcal{N}(0,1)$, $U_i \sim \mathcal{N}(0, 0.2)$ with $\gamma_i \perp\!\!\!\perp U_i$. The parameter δ, which we vary in our simulations, captures the amount of confounding between the intermediate covariate and the outcome. The stronger this confounding, the larger the post-treatment bias for the ACDE when conditioning on Z_i in a naive manner.

The probability that the mediator equals 1 is a function of treatment and all confounders:

$$Pr(M_i = 1|A_i, X_{i1}, X_{i2}, Z_{i1}) = \frac{1}{1 + \exp(-(-2 + X_{i1} - .75X_{i2} + .5A_{i1} + .75Z_{i1}))}$$

To approximate the typical case for matching, where there are many controls to be matched to a smaller number of treated units, this functional form sets the marginal probability of $M_i = 1$ to be

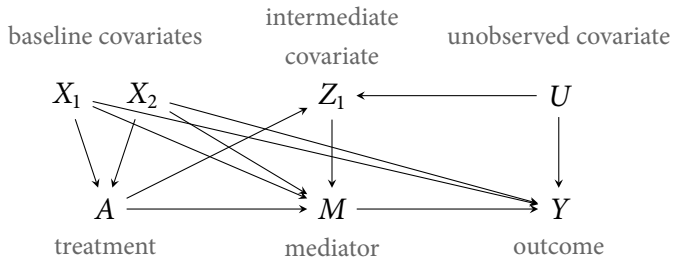

Figure 2: Directed acyclic graph showing the structure of the simulation.

between about 0.24 and 0.28 in our simulations depending on the magnitude of confounding. Finally, the observed outcome Y_i is simulated as

$$Y_i = 210 + 27.4M_i + 13.7X_{i1} + 13.7X_{i2} + \delta U_i + \varepsilon_i$$

where $\varepsilon_i \sim \mathcal{N}(0,1)$ and δ is the same parameter that appears in the functional form of Z_i. In this case, the effect of A_i flows entirely through its effect on the mediator, sp the true controlled direct effect is 0. Figure 2 illustrates the simulation structure in a directed acyclic graph.

As in Kang and Schafer (2007), we simulate model misspecification by considering a scenario where the confounders are not measured directly but rather as the non-linear transformations $(X^*_{i1}, X^*_{i2}, Z^*_i) = \{\exp(X_{i1}/2), (1+\exp(X_{i2}))^{-1} + 10, (Z_i/25 + .6)^3\}$. Were these non-linear transformations known to the researcher, it would be possible to specify the true linear regression model in terms of a correct transformation of the confounders. However, in practice, researchers do not know the exact non-linear transformation that would yield a correctly specified model. Instead, they will typically use models that simply assume linearity and additivity.

Our simulation varies two parameters: sample size and the magnitude of post-treatment confounding (δ). For each simulated dataset, we estimate the controlled direct effect using (1) a naive additive regression estimate that conditions on the mediator and all pre- and post-treatment confounders, (2) a sequential g-estimation approach that assumes the outcome model is linear and additive in all variables, and (3) our telescope matching approach with the Mahalanobis distance metric, bias correcting with the same regression model as in sequential g-estimation. In our simulations, we set the number of units matched to each treated unit to $L = 3$. We also considered an IPTW esti-

mator, and while it performs reasonably well in large samples given correct model specifications for both the mediator and outcome, we omit it from the graphs for expository reasons as the bias under misspecification is far larger than for any of the other methods, consistent with Kang and Schafer (2007). We show the figure with IPTW results in the Supplemental Materials (Figure SM.6).

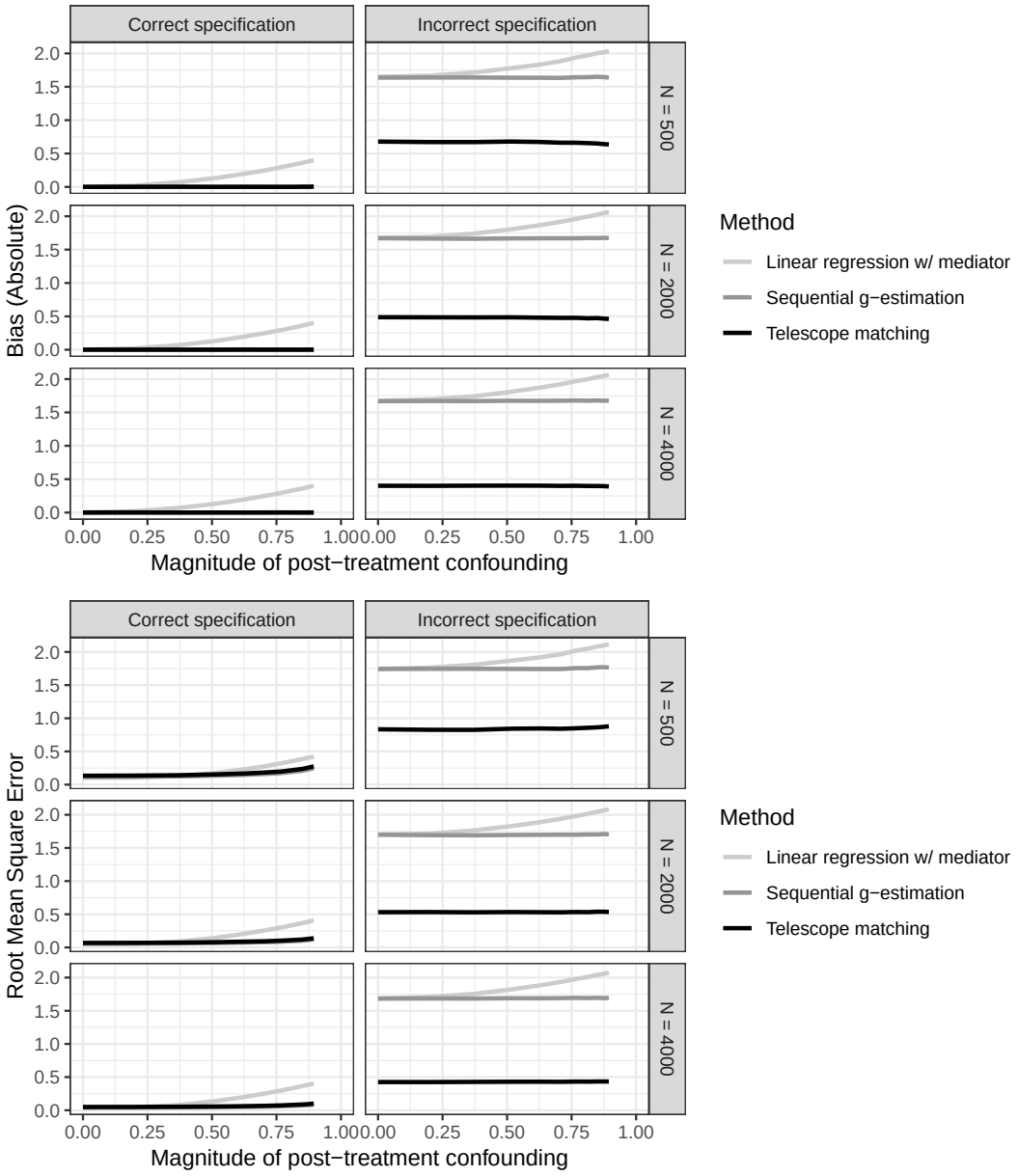

Figure 3: Performance of regression controlling for intermediate covariates, sequential g-estimation, and telescope matching under simulated data with correct and misspecified models. Sequential g-estimation and telescope matching overlap under the correct specification.

Figure 3 plots the absolute value of the bias and the root mean squared error for all three ap-

proaches under correct and incorrect model specifications. On the x-axis, we vary the size of post-treatment confounding, which is measured as the partial correlation between U_i and Y_i. This partial correlation is a deterministic function of δ: $\left(0.2\delta/\sqrt{0.04\delta^2+1}\right)$. For each combination of parameter values, we carried out 10000 iterations of our simulation. We find that both telescope matching and sequential g-estimation are unbiased when the model is correctly specified, with sequential g-estimation having a slight advantage over matching in terms of variance—a gap that decreases significantly as sample size increases. At the larger sample sizes, the increase in variance resulting from including a more flexible imputation model is rather minimal. As expected, simply including all of the covariates in a single regression model suffers from the problem of post-treatment bias, the magnitude of which grows as we increase the correlation between the post-treatment confounder and Y_i. When we introduce model misspecification, the performance of sequential g-estimation is notably worse than that of matching, with the gap growing as a function of the sample size. In particular, we find that for $N = 4000$ the root mean squared error of sequential g-estimation is about four times greater than that of telescope matching when the outcome model is incorrectly specified and inconsistent for the true outcome. Furthermore, the bias and root mean squared error for telescope matching under misspecification are decreasing in N even with a relatively inflexible bias correction model.

We also evaluate the performance of the variance estimators proposed in Section 2.5 for confidence intervals for our matching method. For a fixed level of post-treatment confounding ($\delta = 3.5$), we simulated the coverage rate of 95% confidence intervals calculated using (a) the asymptotic variance estimator defined in (9), (b) the weighted wild bootstrap of Otsu and Rai (2017), and (c) the conventional nonparametric (pairs) bootstrap. Figure 4 plots the coverage rates across different sample sizes.

We find that while the asymptotic and weighted bootstrap variance estimator provide confidence intervals that slightly undercover the true parameter values for smaller sample sizes like $N = 100$, once the sample size is greater than about 500, the coverage rates are generally close to the desired nominal rate, reaching about 94% coverage at $N = 1000$. Indeed, these estimators provide coverage

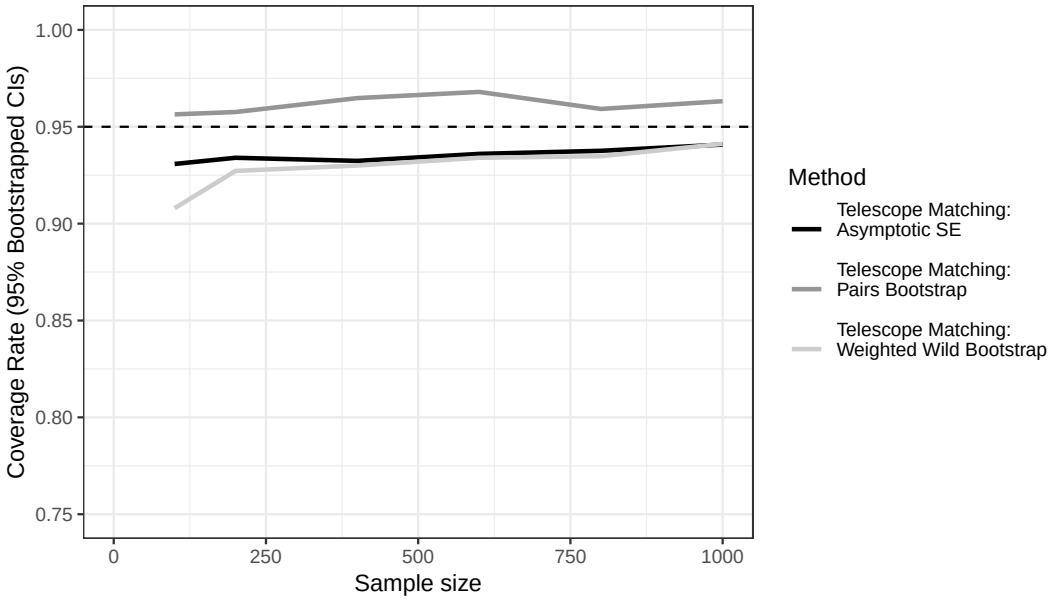

Figure 4: Coverage rate of nominal 95% confidence intervals using, the asymptotic variance estimator (black), the pairs bootstrap estimator (medium gray), and the weighted wild bootstrapped variance estimator (light gray) under correct model specification. Bootstrapped standard errors estimated using 1000 bootstrap iterations. Each coverage rate estimated using 2500 simulation iterations.

similar to the performance of the weighted bootstrap in Otsu and Rai (2017). The standard nonparametric bootstrap on the other hand appears to have slightly higher coverage rates perhaps due to its ability to incorporate the estimation uncertainty in the bias correction, something the analytical and weighted bootstrap variance estimators ignore. Of course, the nonparametric bootstrap is known to have theoretical and practical shortcomings for matching estimators and so we expect its performance to depend heavily on the data generating process. In this case, it appears to over-cover the truth by about 1-2 percentage points. Overall, we find that for reasonable sample sizes, our asymptotic variance estimator and the weighted bootstraping provide a reliable method for constructing confidence intervals when using telescope matching.

Overall, the simulation results are promising for our proposed method. The findings are are consistent with the argument made in Ho et al. (2006) that matching allows researchers to avoid some of the pitfalls of having to choose the "correct" imputation model. Moreover, at least under the data generating process of this simulation, the loss of power when the true model is somehow known is minimal and far outweighed by the reduction in bias under the more likely case where the researcher

happens to select a specification that does not quite match the truth. However, we do caution that, even if it outperforms sequential g-estimation, the bias of telescope matching can still be significant under the incorrect specification without a very large sample.

4 Empirical application: Mental health effects of job training

Huber (2014) examines the effectiveness of IPTW methods for estimating mediation effects in the context of the US Job Corps experiment. This study was conducted during the 1990s to evaluate the impact of a job training program that provided young, low-income Americans with education and technical/vocational instruction. While much of the focus of the evaluation is on job market outcomes, Huber (2014) focuses on potential ancillary effects on participants' quality of life. Specifically, the paper evaluates whether assignment to the job training program impacts participants' self-reported health and the extent to which the driving mechanism is the effect of training on subsequent employment. In this application, the treatment of interest A_i is an indicator denoting whether a participant was randomly assigned to be eligible for the Job Corps program, the mediator M_i is also an indicator denoting whether the participant was employed in the first half of the second year post-randomization. We follow Huber (2014) and define the outcome Y_i as an indicator for whether the participant reported "very good" general health in a follow-up survey conducted 2.5 years after randomization. Both pre-treatment covariates X_i and intermediate covariates Z_i that are measured slightly prior to the mediator are observed for each participant.

Huber (2014) recognizes that standard identification strategies for mediation quantities assume sequential ignorability of the mediator conditional only on pre-treatment covariates and that this assumption may be implausible in the job training context. Other factors, in particular whether individuals actually participated in job training, that are likely affected by assignment to treatment may infuence both subsequent employment and health. Thus, Huber (2014) weakens the standard mediation assumptions to allow for intermediate confounders, as we do above, but uses this to identify and estimate a type of natural direct effect that isolates the pathways from treatment to outcome ex-

cluding those that include the mediator or any intermediate confounders. We instead focus here on the controlled direct effect of treatment fixing the mediator, but allowing for effects through intermediate confounders. We believe this helps to better isolate the causal mechanisms at work since we are only "blocking" the role of one variable—the mediator.

We estimate the controlled direct effects of assignment to job training on reporting "very good" health after 2.5 years out fixing employment measured in the period 1–1.5 years after assignment to either "employed" or "unemployed." We use several techniques to estimate this quantity: (1) an "over-controlled" regression that includes all baseline and intermediate covariates; (2) sequential g-estimation assuming linear additive models for the covariates; and (3) telescope matching with 1-to-3 matching. Following, Huber (2014) we analyze male and female particpants separately as the authors suspect that program effects are heterogeneous by gender. We document the specific covariates used in the Supplemental Material.

Figure 5 plots the estimated ACDEs of assignment to job training on health among women and men fixing intermediate employment to either "employed" or "unemployed." First, we replicate the main average treatment effect: among women, assignment to job training improves the probability of self-reporting very good health by about 2.5 percentage points ($p < .1$). We find that the ACDE estimates from the sequential g-estimation approach generally replicate the natural direct effect estimates from Huber (2014) in terms of direction and statistical significance. There appears to be a positive effect of training on health for women that does not flow through its effects on employment. On average, women assigned to receive job training are 4 percentage points more likely to report very good health, fixing employment status to "employed." In this case, telescope matching yields a similar estimate to sequential g-estimation, though the matching estimate is slightly larger and has a 90% confidence interval that does not cover 0. However, telescope matching and sequential g-estimation diverge substantially when considering the ACDE setting employment to "unemployed." Sequential g-estimation suggests no direct effect of training on health when participating women are unemployed while telescope matching suggests that there are potential *harmful* effects of assignment to training on the health of women who are unemployed after one year. Women assigned to the pro-

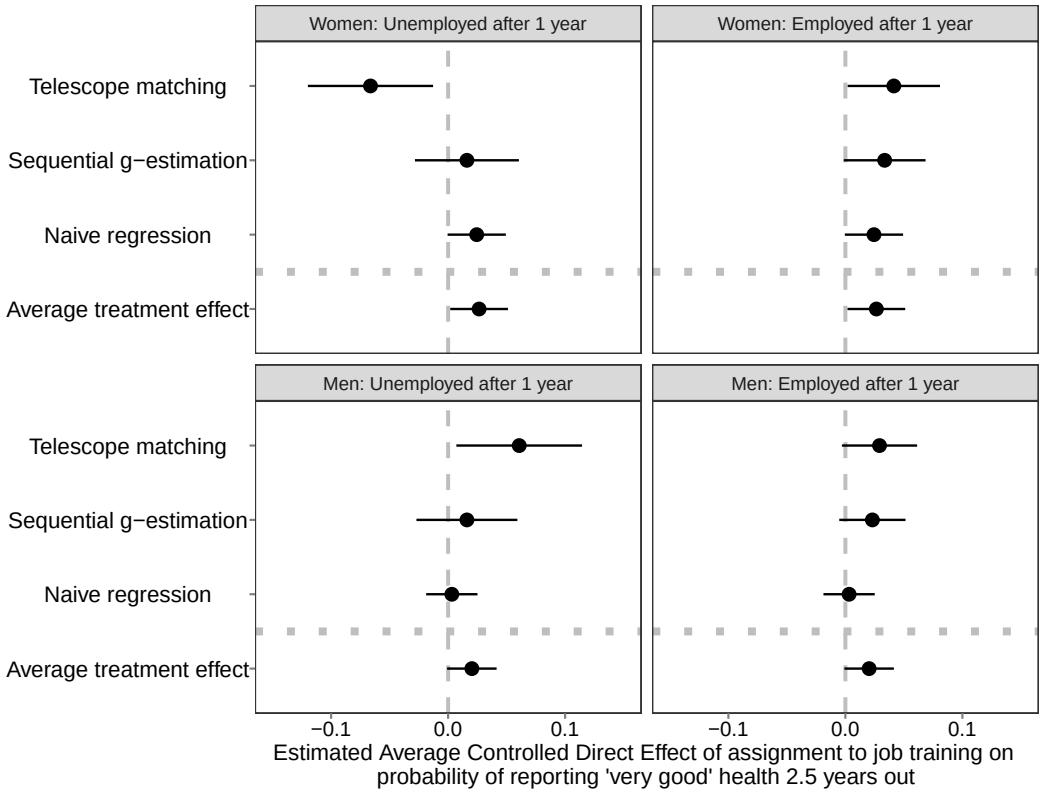

Figure 5: Estimated ACDEs of assignment to training among women (N = 4352) and men (N = 5673). Lines denote 90% confidence intervals. Sequential g-estimation CIs estimated using 1000 bootstrap iterations. Telescope matching CIs estimated using asymptotic variance estimator.

gram are about 7 percentage points less likely to report very good health when unemployed after 1 year. This suggests the possibility of significant treatment effect heterogeneity in the downstream effects of training resulting from an interaction with participants' labor market outcomes.

We see the opposite pattern among men. Sequential-g estimation and telescope matching recover slightly positive but statistically insignificant direct effects of training among men who are employed. But the direct effect among men assigned to "unemployment" post-treatment suggests training has a slight positive impact on self-assessed health (about 6 percentage points). In the Supplementary Materials, we discuss in greater detail the likely source of misspecification that explains the discrepancy between sequential g-estimation and telescope matching. Specifically, we find the presence of higher-order interactions between pre- and post-treatment employment status and outcome that are not included in the parsimonious additive models.

Overall, adopting the less parametric telescope matching approach reveals some novel insights regarding the direct consequences of job training interventions on health outcomes. In particular, we find that employment outcomes do not fully account for all the possible mechanisms by which training influences participants' health. Moreover, the impact of training while holding fixed employment is not uniformly positive, especially for women. We find that while those women assigned to the program exhibited slightly higher self-reported health when employed post-treatment, the causal effect of training actually appears to be *negative* among those who are unemployed after one year. Consistent with the original conclusions of Huber (2014), the training-employment-health pathway does not explain away the effect of the program, but employment may nevertheless act as a post-treatment *moderator*.

5 Conclusion

In this paper, we have introduced a novel method for estimating the direct effect of treatment for fixed values of a mediator. This matching-based approach flexibly imputes missing values of the potential outcomes and appears to be more robust to model misspecification than other approaches like sequential g-estimation. This method could be of use to many applied researchers who want to estimate direct effects but have a large degree of uncertainty about the correct model specification for baseline and intermediate covariates. Furthermore, we derived several properties of the estimator, including its large-sample distribution, that allowed us to develop a bias-corrected version of this estimator that augments the matching with regression.

There are several avenues for future work on this frontier. First, it would be interesting to understand how these methods could be extended to estimate quantities of interest in mediation analyses like the natural direct and indirect effect, when the assumptions of that setting holds. Second, we have explored bias correction through simple additive linear regression models but a range of more flexible regression techniques, from generalized additive models to cutting-edge machine learning methods, could plausibly be used as well. In general, this paper illustrates how estimation of con-

trolled direct effects can be treated as a problem of imputing missing potential outcomes $Y_i(a, 0)$. We outline one particular imputation strategy, a two-stage matching estimator, but there are many other imputation methods, each with their own particular advantages and drawbacks, that could be investigated in subsequent research.

Bibliography

Abadie, Alberto, and Guido W. Imbens. 2006. "Large sample properties of matching estimators for average treatment effects." *Econometrica* 74 (1): 235–267.

Abadie, Alberto, and Guido W. Imbens. 2008. "On the Failure of the Bootstrap for Matching Estimators." *Econometrica* 76 (6): 1537–1557.

Abadie, Alberto, and Guido W. Imbens. 2011. "Bias-Corrected Matching Estimators for Average Treatment Effects." *Journal of Business & Economic Statistics* 29 (January): 1–11.

Abadie, Alberto, and Guido W Imbens. 2012. "A martingale representation for matching estimators." *Journal of the American Statistical Association* 107 (498): 833–843.

Abadie, Alberto, and Guido W. Imbens. 2016. "Matching on the Estimated Propensity Score." *Econometrica* 84 (2): 781–807.

Acharya, Avidit, Matthew Blackwell, and Maya Sen. 2016. "Explaining Causal Findings Without Bias: Detecting and Assessing Direct Effects." *American Political Science Review* 110 (3): 512–529.

Acharya, Avidit, Matthew Blackwell, and Maya Sen. 2018. "Analyzing Causal Mechanisms in Survey Experiments." *Political Analysis* 26 (4): 357–378.

Billingsley, Patrick. 1995. *Probability and Measure*. 3 ed. New York: John Wiley and Sons.

Blackwell, Matthew, and Adam Glynn. 2018. "How to Make Causal Inferences with Time-Series Cross-Sectional Data under Selection on Observables." *American Political Science Review* 112 (4): 1067–1082.

Dehejia, Rajeev H., and Sadek Wahba. 1999. "Causal Effects in Nonexperimental Studies: Reevaluating the Evaluation of Training Programs." *Journal of the American Statistical Association* 94 (448): 1053–1062.

Ho, Daniel E., Kosuke Imai, Gary King, and Elizabeth A. Stuart. 2006. "Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference." *Political Analysis* 15 (3): 199.

Huber, Martin. 2014. "Identifying causal mechanisms (primarily) based on inverse probability weighting." *Journal of Applied Econometrics* 29 (6): 920–943.

Huber, Martin, Michael Lechner, and Anthony Strittmatter. 2018. "Direct and indirect effects of training vouchers for the unemployed." *Journal of the Royal Statistical Society. Series A (Statistics in Society)* 181 (2): 441–463.

Imai, Kosuke, Luke Keele, and Teppei Yamamoto. 2010. "Identification, Inference and Sensitivity Analysis for Causal Mediation Effects." *Statistical Science* 25 (1): 51–71.

Imbens, Guido W. 2004. "Nonparametric Estimation of Average Treatment Effects Under Exogeneity: A Review." *Review of Economics and Statistics* 86 (1): 4–29.

Kang, Joseph D.Y., and Joseph L. Schafer. 2007. "Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data." *Statistical science* 22 (4): 523–539.

Lechner, Michael, and Ruth Miquel. 2010. "Identification of the effects of dynamic treatments by sequential conditional independence assumptions." *Empirical Economics* 39 (August): 111–137.

Mammen, Enno. 1993. "Bootstrap and Wild Bootstrap for High Dimensional Linear Models." *The Annals of Statistics* 21 (1): 255–285.

Otsu, Taisuke, and Yoshiyasu Rai. 2017. "Bootstrap inference of matching estimators for average treatment effects." *Journal of the American Statistical Association* 112 (520): 1720–1732.

Richardson, Thomas S., and Andrea Rotnitzky. 2014. "Causal Etiology of the Research of James M. Robins." *Statistical Science* 29 (11): 459–484.

Robins, James M. 1986. "A new approach to causal inference in mortality studies with sustained exposure periods-Application to control of the healthy worker survivor effect." *Mathematical Modelling* 7 (9-12): 1393-1512.

Robins, James M. 1997. "Causal Inference from Complex Longitudinal Data." In *Latent Variable Modeling and Applications to Causality*, ed. M. Berkane. Vol. 120 of *Lecture Notes in Statistics* New York: Springer-Verlag.

Robins, James M. 2000. "Marginal Structural Models versus Structural Nested Models as Tools for Causal Inference." In *Statistical Models in Epidemiology, the Environment, and Clinical Trials*, ed. M. Elizabeth Halloran and Donald Berry. Vol. 116 of *The IMA Volumes in Mathematics and its Applications* New York: Springer-Verlag.

Robins, James M., and Andrea Rotnitzky. 2004. "Estimation of treatment effects in randomised trials with non-compliance and a dichotomous outcome using structural mean models." *Biometrika* 91 (4): 763-783.

Robins, James M., and Sander Greenland. 1992. "Identifiability and Exchangeability for Direct and Indirect Effects." *Epidemiology* 3 (2): 143–155.

Rosenbaum, Paul R. 1984. "The consquences of adjustment for a concomitant variable that has been affected by the treatment." *Journal of the Royal Statistical Society. Series A (General)* 147 (5): 656–666.

Rosenbaum, Paul R. 1995. *Observational Studies*. New York: Springer-Verlag.

Rubin, Donald B. 1974. "Estimating causal effects of treatments in randomized and nonrandomized studies." *Journal of educational Psychology* 66 (5): 688.

VanderWeele, Tyler. 2015. *Explanation in causal inference: methods for mediation and interaction*. Oxford: Oxford University Press.

Supplemental Materials

A Proofs

We begin by showing the decomposition of the simple telescope matching estimator. Based on Abadie and Imbens (2006), we can write the telescope matching estimator in the following linear form:

$$\widehat{\tau} = \frac{1}{N}\sum_{i=1}^{N}(2A_i - 1)\left(1 + \frac{K_L^a(i)}{L}\right)\widehat{Y}_i(A_i, 0)$$

Two identities that will greatly simplify our derivations. For any variable W_i, we have the following:

$$\frac{1}{N}\sum_{i=1}^{N}(2A_i - 1)M_i\frac{K_L^m(i)}{L}W_i = \frac{1}{N}\sum_{i=1}^{N}(2A_i - 1)(1 - M_i)\left(\frac{1}{L}\sum_{\ell\in\mathcal{J}_L^m(i)}W_\ell\right) \tag{11}$$

$$\frac{1}{N}\sum_{i=1}^{N}(2A_i - 1)M_i\frac{K_L^{am}(i)}{L^2}W_i = \frac{1}{N}\sum_{i=1}^{N}(2A_i - 1)(1 - M_i)\frac{K_L^a(i)}{L}\left(\frac{1}{L}\sum_{\ell\in\mathcal{J}_L^m(i)}W_\ell\right) \tag{12}$$

$$\frac{1}{N}\sum_{i=1}^{N}(2A_i - 1)\frac{K_L^a(i)}{L}W_i = -\frac{1}{N}\sum_{i}(2A_i - 1)\left(\frac{1}{L}\sum_{j\in\mathcal{J}_L^a(i)}W_j\right) \tag{13}$$

Replacing $\widehat{Y}_i(A_i, 0)$ with its definition and applying (11) and (12) gives:

$$\begin{aligned}
\widehat{\tau} &= \frac{1}{N}\sum_{i=1}^{N}(2A_i - 1)\left(1 + \frac{K_L^a(i)}{L}\right)(1 - M_i)Y_i + \frac{1}{N}\sum_{i=1}^{N}(2A_i - 1)M_i\left(\frac{1}{L}\sum_{\ell\in\mathcal{J}_L^m(i)}Y_\ell\right) \\
&\quad + \frac{1}{N}\sum_{i=1}^{N}(2A_i - 1)\frac{K_L^a(i)}{L}M_i\left(\frac{1}{L}\sum_{\ell\in\mathcal{J}_L^m(i)}Y_\ell\right) \\
&= \frac{1}{N}\sum_{i=1}^{N}(2A_i - 1)(1 - M_i)\left(1 + \frac{K_L^a(i)}{L} + \frac{K_L^m(i)}{L} + \frac{K_L^{am}(i)}{L^2}\right)Y_i
\end{aligned}$$

Note that $Y_i = \mu_{A_i 0}(X_i, A_i) + \varepsilon_i + \eta_i$. Based on this, we can write this as the following:

$$\widehat{\tau} = E_L^m \tag{14}$$

$$+ \frac{1}{N}\sum_{i=1}^{N}(2A_i - 1)\left(1 + \frac{K_L^a(i)}{L} + \frac{K_L^m(i)}{L} + \frac{K_L^{am}(i)}{L^2}\right)(1 - M_i)\eta_i \tag{15}$$

$$+ \frac{1}{N}\sum_{i=1}^{N}(2A_i - 1)\left(1 + \frac{K_L^a(i)}{L} + \frac{K_L^m(i)}{L} + \frac{K_L^{am}(i)}{L^2}\right)(1 - M_i)\mu_{A_i 0}(X_i, A_i) \tag{16}$$

Rearranging terms, we have:

$$\widehat{\tau} = E_L^m + E_L^a + \frac{1}{N}\sum_{i=1}^{N}(2A_i - 1)\left(1 + \frac{K_L^a(i)}{L}\right)\mu_{A_i 0}(X_i, A_i) \tag{17}$$

$$+ \frac{1}{N}\sum_{i=1}^{N}(2A_i - 1)(1 - M_i)\left(\frac{K_L^m(i)}{L}\right)(1 - M_i)\mu_{A_i 0}(X_i, Z_i, A_i) \tag{18}$$

$$+ \frac{1}{N}\sum_{i=1}^{N}(2A_i - 1)(1 - M_i)\left(\frac{K_L^{am}(i)}{L^2}\right)(1 - M_i)\mu_{A_i 0}(X_i, Z_i, A_i) \tag{19}$$

$$- \frac{1}{N}\sum_{i=1}^{N}(2A_i - 1)M_i\mu_{A_i 0}(X_i, Z_i, A_i) - \frac{1}{N}\sum_{i=1}^{N}(2A_i - 1)M_i\frac{K_L^a(i)}{L}\mu_{A_i 0}(X_i, Z_i, A_i) \tag{20}$$

Using (11) and 12 on the last three lines of this expression and combining terms gives:

$$\widehat{\tau} = E_L^m + E_L^a + \frac{1}{N}\sum_{i=1}^{N}(2A_i - 1)\left(1 + \frac{K_L^a(i)}{L}\right)\mu_{A_i 0}(X_i, A_i) \tag{21}$$

$$+ \frac{1}{N}\sum_{i=1}^{N}(2A_i - 1)M_i\left(1 + \frac{K_L^a(i)}{L}\right)\left(\frac{1}{L}\sum_{\ell \in \mathcal{J}_L^m(i)}\mu_{A_i 0}(X_\ell, Z_\ell, A_i) - \mu_{A_i 0}(X_i, Z_i, A_i)\right) \tag{22}$$

$$= E_L^m + E_L^a + B_L^m + \frac{1}{N}\sum_{i=1}^{N}(2A_i - 1)\left(1 + \frac{K_L^a(i)}{L}\right)\mu_{A_i 0}(X_i, A_i) \tag{23}$$

Finally, adding and subtracting $(1/N)\sum_{i=1}^{N}(2A_i - 1)\mu_{1-A_i,0}(X_i, 1 - A_i)$ gives the final decomposition:

$$\widehat{\tau} = \frac{1}{N}\sum_{i=1}^{N}\tau(X_i) + E_L^m + E_L^a + B_L^m + B_L^a. \tag{24}$$

The following lemma ensures that the implied weights of the matching estimator are uniformly bounded. Its proof is rather long and involved and largely based on a similar proof in Abadie and Imbens (2006), so we omit it.

Lemma 1. *Suppose that Assumptions 1, 2, and 3 hold. Then, (i) the expectation* $\mathbb{E}[(K_L^{am}(i))^q]$ *is uniformly bounded in N.*

Proof of Lemma 1. Proof available on the author's website. □

Proof of Theorem 1. Let $D_N = \frac{1}{N}\sum_{i=1}^{N}\tau(X_i) - \tau + E_L^a + E_L^m$. We can write

$$\sqrt{N}D_N = \sum_{k=1}^{3N}\xi_{N,k},$$

where

$$\xi_{N,k} = \begin{cases} \frac{1}{\sqrt{N}}\left(\tau(X_i) - \tau\right), & \text{if } 1 \le k \le N \\ \frac{1}{\sqrt{N}}(2A_i - 1)\left(1 + \frac{K_L^a(i)}{L}\right)\eta_i, & \text{if } N+1 \le k \le 2N \\ \frac{1}{\sqrt{N}}(2A_i - 1)(1 - M_i)\left(1 + \frac{K_L^a(i)}{L} + \frac{K_L^m(i)}{L} + \frac{K_L^{am}(i)}{L^2}\right)\varepsilon_i & \text{if } 2N+1 \le k \le 3N \end{cases}$$

Let $\mathbf{X} = \{X_1, \ldots, X_N\}$, $\mathbf{A} = \{A_1, \ldots, A_N\}$, $\mathbf{Z} = \{Z_1, \ldots, Z_N\}$, and $\mathcal{M}_N = \{M_1, \ldots, M_N\}$. We then define the following σ-fields:

$$\mathcal{F}_{N,k} = \begin{cases} \sigma\{\mathbf{A}, X_1, \ldots, X_k\} & \text{for } 1 \le k \le N \\ \sigma\{\mathbf{A}, \mathbf{X}, Z_1, \ldots, Z_{k-N}\} & \text{for } N+1 \le k \le 2N \\ \sigma\{\mathbf{A}, \mathbf{X}, \mathbf{Z}, \mathbf{M}, Y_1, \ldots, Y_{k-2N}\} & \text{for } 2N+1 \le k \le 3N \end{cases}$$

Following the logic of Abadie and Imbens (2012), we note that

$$\left\{\sum_{j=1}^{i}\xi_{N,j}, \mathcal{F}_{N,i}, 1 \le i \le 3N\right\}$$

is a martingale for $N \ge 1$.

For $1 \le k \le N$, the conditional variances of the martingale differences are given by

$$\begin{aligned}\mathbb{E}[\xi_{N,k}^2|\mathcal{F}_{N,k-1}] &= \frac{1}{N}\mathbb{E}[(\tau(X_i) - \tau)^2 | A_i], \\ &= \frac{1}{N}\mathbb{E}[(\tau(X_i) - \tau)^2],\end{aligned}$$

where the second equality holds by ignorability. For $N+1 \leq k \leq 2N$, the conditional variances are

$$\begin{aligned}\mathbb{E}[\xi_{N,k}^2|\mathcal{F}_{N,k-1}] &= \frac{1}{N}\left(1+\frac{K_L^a(i)}{L}\right)^2 \mathbb{E}[\eta_i^2|\mathbf{X},\mathbf{A}], \\ &= \frac{1}{N}\left(1+\frac{K_L^a(i)}{L}\right)^2 \sigma_\eta^2(X_i,A_i)\end{aligned}$$

Finally, for the $2N+1 \leq k \leq 3N$, the conditional variances of the martingale differences are

$$\begin{aligned}\mathbb{E}[\xi_{N,k}^2|\mathcal{F}_{N,k-1}] &= \frac{1}{N}(1-M_i)\left(1+\frac{K_L^a(i)}{L}+\frac{K_L^m(i)}{L}+\frac{K_L^{am}(i)}{L^2}\right)\mathbb{E}[\varepsilon_i^2|\mathbf{X},\mathbf{A},\mathbf{Z},\mathbf{M}], \\ &= \frac{1}{N}(1-M_i)\left(1+\frac{K_L^a(i)}{L}+\frac{K_L^m(i)}{L}+\frac{K_L^{am}(i)}{L^2}\right)\sigma^2(X_i,Z_i,A_i,0).\end{aligned}$$

Thus, we can invoke a weak law of large numbers argument to show that

$$\sum_{k=1}^{3N}\mathbb{E}[\xi_{N,k}^2|\mathcal{F}_{N,k-1}] \xrightarrow{p} \sigma^2,$$

where

$$\begin{aligned}\sigma^2 &= \mathbb{E}[(\tau(X_i)-\tau)^2] \\ &+ \mathbb{E}\left[\left(1+\frac{K_L^a(i)}{L}\right)^2\sigma_\eta^2(X_i,A_i)\right] \\ &+ \mathbb{E}\left[(1-M_i)\left(1+\frac{K_L^a(i)}{L}+\frac{K_L^m(i)}{L}+\frac{K_L^{am}(i)}{L^2}\right)^2\sigma^2(X_i,Z_i,A_i,0)\right]\end{aligned} \tag{25}$$

To establish the large sample distribution of D_N, we use a Lyapunov condition:

$$\sum_{k=1}^{3N}\mathbb{E}[|\xi_{N,k}|^{2+\delta}] \to 0, \qquad \text{for some } \delta > 0.$$

Note that this condition implies the more standard Lindeberg condition used in martingal central limit theorems (Billingsley, 1995; Abadie and Imbens, 2012).

From Lemma 3 of Abadie and Imbens (2006) and Lemma 1, we have that $E[(K_L^a(i)/L)^4]$, $E[(1+K_L^m(i)/L)^4]$, and $\mathbb{E}[(K_L^{am}(i))^4]$ are uniformly bounded. Through iterated use of Minkowski inequality, we have

$$\mathbb{E}\left[\left(1+\frac{K_L^a(i)}{L}+\frac{K_L^m(i)}{L}+\frac{K_L^{am}(i)}{L^2}\right)^4\right] \leq \mathbb{E}\left[\left(1+\frac{K_L^m(i)}{L}\right)^4\right]^{1/4}+\left(E\left[\left(\frac{K_L^a(i)}{L}\right)^4\right]^{1/4}+E\left[\left(\frac{K_L^{am}(i)}{L^2}\right)^4\right]^{1/4}\right)^{1/4}$$

which implies that $(1+K_L^a/L+K_L^m/L+K_L^{am}(i)/L^2)$ are uniformly bounded.

Letting $\delta = 2$, look at one term from $2N+1 \leq k \leq 3N$:

$$\mathbb{E}[\xi_{N,k}^4] = \frac{1}{N^2}\mathbb{E}\left[(1-M_i)\left(1+\frac{K_L^a(i)}{L}+\frac{K_L^m(i)}{L}+\frac{K_L^{am}(i)}{L^2}\right)^4 \mathbb{E}[\varepsilon_i^4|X_iA_i,Z_i,M_i]\right].$$

The expectation on the right-hand side is bounded given the bound on fourth moments for Y_i (part (iv) of Assumption 3) and the earlier mentioned uniform bound on $(1+K_L^a/L+K_L^m/L+K_L^{am}(i)/L^2)$. Similar analyses can be conducted for the other two parts of the martingale, which will ensure the Lyapunov condition holds. Thus, based on the martingale central limit theorem (Billingsley, 1995, Theorem 35.12), we have $\sqrt{N}D_N \xrightarrow{d} N(0,\sigma^2)$.

□

Proof of Theorem 2. Define the following:

$$\widetilde{B}_L^a = \frac{1}{N}\sum_{i=1}^N (2A_i-1)\left[\frac{1}{L}\sum_{j\in J_L^a(i)} \widetilde{\mu}_{1-A_i,0}(X_i,1-A_i) - \widetilde{\mu}_{1-A_i,0}(X_j,1-A_i)\right]$$

where, $\widetilde{\mu}_{a0}(x,a) = \mathbb{E}[\widetilde{Y}_{i0}|X_i=x,A_i=a]$.

Let Λ_ℓ be the set of vectors λ such that $|\lambda| = \ell$ and let $\partial^\lambda g(x) = \partial^{|\lambda|}g(x)/\partial x_1^{\lambda_1}\cdots\partial x_k^{\lambda_k}$. Finally, for $d \geq 0$, define $|g|_d = \max_{|\lambda|\leq d}\sup_x|\partial^\lambda g(x)|$. Lemma A.1 of Abadie and Imbens (2011) uniformly bounded discrepancy between the partial derivatives of the series estimator and the true CEF, which here implies that:

$$|\widehat{\mu}_{a0}-\mu_{a0}|_d = \max_{|\lambda|\leq d}\sup_{(x,z)\in\mathbb{V}}|\partial^\lambda(\widehat{\mu}_{a0}(x,z)-\mu_{a0}(x,z))| = O_p(G^{1+2d}((G/N)^{1/2}+G^{-\zeta})). \tag{26}$$

This convergence ensures that $|\widehat{B}_L^m - B_L^m| = o_p(N^{-1/2})$ by the same logic as the proofs in Lemma A.2 and Theorem 2 of Abadie and Imbens (2011). Similar logic will ensure that $|\widehat{B}_L^a - \widetilde{B}_L^a| = o_p(N^{-1/2})$. By the triangle inequality, we have $|\widehat{B}_L^a - B_L^a| \leq |\widehat{B}_L^a - \widetilde{B}_L^a| + |\widetilde{B}_L^a - B_L^a|$.

We could apply the same logic as these other steps to $|\widetilde{B}_L^a - B_L^a|$ if we had a similar uniform convergence result for $|\widetilde{\mu}_a - \mu_a|_d$. To do so, we use the definition of $\widetilde{Y}_{i0}$ to derive the following:

$$\widetilde{\mu}_{a0}(x) = \mu_{a0}(x) + b(x,a)$$

where,

$$b_a(x) = \mathbb{E}\left[M_i \left(\frac{1}{L} \sum_{\ell \in J_L^m(i)} \widehat{\mu}(x, a, Z_i, 0) - \widehat{\mu}(X_\ell, a, Z_\ell, 0) - (\mu(x, a, Z_i, 0) - \mu(X_\ell, a, Z_\ell, 0)) \right) \middle| X_i = x, A_i = a \right]$$

Thus, we have: $|\widetilde{\mu}_{a0} - \mu_{a0}|_d = \max_{|\lambda| \leq d} \sup_{x \in \mathbb{X}} |\partial^\lambda b_a(x)|$. Note that in this second stage, we only consider partial derivatives with respect to the baseline covariates, where as in the first stage, we consider partial derivatives with respect to the baseline and intermediate covariates. Thus, for the functions, $\mu_{a0}(x, z)$ and its estimate, $\widehat{\mu}_{a0}(x, z)$, the above derivative discrepancy norm with respect to just x will be bounded by the one that covers x and z in (26). Thus, the interior of expectation of the $b_a(x)$ function can be bounded and is integrable, which means we can use Lebesgue's dominated convergence theorem to switch the order of differentiation and expectation (while also applying the triangle inequality):

$$|\partial^\lambda b_a(x)| \leq \mathbb{E}\left[\frac{1}{L} \sum_{\ell \in J_L^m(i)} \left| \partial^\lambda \left(\widehat{\mu}(x, a, Z_i, 0) - \widehat{\mu}(X_\ell, a, Z_\ell, 0) - (\mu(x, a, Z_i, 0) - \mu(X_\ell, a, Z_\ell, 0)) \right) \right| \middle| X_i = x, A_i = a \right]$$

Again, because the difference in the derivatives of these expectations are bounded, the entire function inside the expectation will be bounded by $2 \times |\widehat{\mu}_{a0} - \mu_{a0}|_d$. Thus, we have:

$$|\widetilde{\mu}_{a0} - \mu_{a0}|_d = O_p(G^{1+2d}((G/N)^{1/2} + G^{-\zeta})). \tag{27}$$

Using the same logic as Lemma A.2 in Abadie and Imbens (2011), this allows us to show that

$$\max_{i=1,\ldots,N} |\widetilde{\mu}_{a0}(X_i) - \widetilde{\mu}_{a0}(X_\ell) - (\mu_{a0}(X_i) - \mu_{a0}(X_\ell))| = o_p(N^{-1/2})$$

for $a = 0, 1$. We can then apply Theorem 2 of Abadie and Imbens (2011) to show $|\widehat{B}_L^a - B_L^a| = o_p(N^{-1/2})$, thus showing that the bias correction will not affect the asymptotic variance of the matching estimator.

□

Proof of Theorem 3. We first show that $|\widehat{V}^\eta - V^\eta| = o_p(1)$. Let $\widehat{u}_{ia} = \widehat{\mu}(X_i, Z_i, A_i, 0) - \widehat{\mu}_{A_i0}(X_i, A_i)$ be the "residuals" from the difference between the first-stage and second-stage CEFs and let $u_{ia} = \mu(X_i, Z_i, A_i, 0) - \mu_{A_i0}(X_i, A_i)$ be the corresponding population error. This allows us to write $\widehat{V}^\eta =$

$(1/N)\sum_{i=1}^{N}(1+K_L^a(i)/L)^2\widehat{u}_{ia}^2$ and $\sigma_\eta^2(X_i,A_i)=\mathbb{E}[u_{ia}^2|X_i,A_i]$. Let $\delta_{im}=\widehat{\mu}(X_i,Z_i,A_i,0)-\mu(X_i,Z_i,A_i,0)$ and $\delta_{ia}=\left(\widehat{\mu}_{A_i0}(X_i,A_i)-\mu_{A_i0}(X_i,A_i)\right)$. Then we have

$$\widehat{u}_{ia}=u_{ia}+\delta_{ia}+\delta_{im}.$$

To show the result, it is sufficient to show that

$$\frac{1}{N}\sum_{i=1}(K_L^a(i))^q\left(\widehat{u}_{ia}^2-\sigma_\eta^2(X_i,A_i)\right)=o_p(1).$$

We can write this quantity as

$$\begin{aligned}\frac{1}{N}\sum_{i=1}(K_L^a(i))^q\left(\widehat{u}_{ia}^2-\sigma_\eta^2(X_i,A_i)\right)&=\frac{1}{N}\sum_{i=1}(K_L^a(i))^q\left(u_{ia}^2-\mathbb{E}[u_{ia}^2|X_i,A_i]\right)\\&+\frac{1}{N}\sum_{i=1}(K_L^a(i))^q\delta_{ia}^2+\frac{1}{N}\sum_{i=1}(K_L^a(i))^q\delta_{im}^2+\frac{2}{N}\sum_{i=1}(K_L^a(i))^q\left(\delta_{ia}\delta_{im}\right)\\&+\frac{2}{N}\sum_{i=1}(K_L^a(i))^q\left(u_{ia}\delta_{ia}\right)+\frac{2}{N}\sum_{i=1}(K_L^a(i))^q\left(u_{ia}\delta_{im}\right)\end{aligned}\tag{28}$$

Based on (26), (27), and Assumption 4, the maximal estimation errors:

$$\overline{\delta}_a=\max_{i=1,\ldots,N}|\delta_{ia}|=o_p(1)\qquad\overline{\delta}_m=\max_{i=1,\ldots,N}|\delta_{im}|=o_p(1).$$

Furthermore, both $(K_L^a(i))^q$ and u_{ia} have bounded moments. With these results, it is easy to show that all of the terms on the right-hand side of (28) will be $o_p(1)$ except for the first, leaving:

$$\frac{1}{N}\sum_{i=1}(K_L^a(i))^q\left(\widehat{u}_{ia}^2-\sigma_\eta^2(X_i,A_i)\right)=\frac{1}{N}\sum_{i=1}(K_L^a(i))^q\left(u_{ia}^2-\mathbb{E}[u_{ia}^2|X_i,A_i]\right)+o_p(1)\tag{29}$$

Finally, the law of large numbers will ensure that $(1/N)\sum_{i=1}(K_L^a(i))^q u_{ia}^2$ and $(1/N)\sum_{i=1}(K_L^a(i))\mathbb{E}[u_{ia}^2|X_i,A_i]$ to the same value, resulting in:

$$\frac{1}{N}\sum_{i=1}(K_L^a(i))^q\left(\widehat{u}_{ia}^2-\sigma_\eta^2(X_i,A_i)\right)=o_p(1).\tag{30}$$

This implies $|\widehat{V}^\eta-V^\eta|=o_p(1)$. A similar derivation shows that $|\widehat{V}^\varepsilon-V^\varepsilon|=o_p(1)$.

Next we show that $|\widehat{V}^{\tau(X)}-V^{\tau(X)}|=o_p(1)$. Let $\delta_{i1}=\widehat{\mu}_{10}(X_i)-\mu_{10}(X_i)$ and $\delta_{i0}=\widehat{\mu}_{00}(X_i)-\mu_{00}(X_i)$. We can decompose the model-predicted effects as:

$$\widehat{\mu}_{10}(X_i)-\widehat{\mu}_{00}(X_i)=\tau(X_i)+\delta_{i1}-\delta_{i0}$$

Let $\overline{\tau} = \widehat{\mu}_{10}(X_i) - \widehat{\mu}_{00}(X_i)$. Note that $\overline{\tau} \xrightarrow{p} \tau$, given the consistency of the series estimators. This implies that $|\overline{\tau} - \widetilde{\tau}| = o_p(1)$. Thus, we can write:

$$\begin{aligned}
\widehat{V}^{\tau(X)} &= \frac{1}{N}\sum_{i=1}^{N} \left(\widehat{\mu}_{10}(X_i) - \widehat{\mu}_{00}(X_i) - \overline{\tau}\right)^2 - \left(\overline{\tau} - \widetilde{\tau}\right)^2 \\
&= \frac{1}{N}\sum_{i=1}^{N} \left(\widehat{\mu}_{10}(X_i) - \widehat{\mu}_{00}(X_i) - \tau\right)^2 - \left(\overline{\tau} - \tau\right)^2 + o_p(1) \\
&= \frac{1}{N}\sum_{i=1}^{N} \left(\widehat{\mu}_{10}(X_i) - \widehat{\mu}_{00}(X_i) - \tau\right)^2 + o_p(1) \\
&= \frac{1}{N}\sum_{i=1}^{N} \left(\tau(X_i) - \tau + \delta_{i1} - \delta_{i0}\right)^2 + o_p(1)
\end{aligned}$$

As above, the estimation errors here, δ_{i1} and δ_{i0} are bounded by terms that converge to o. Furthermore, $(1/N)\sum_{i=1}^{N}(\tau(X_i) - \tau) = o_p(1)$. These two facts imply that

$$\widehat{V}^{\tau(X)} = \frac{1}{N}\sum_{i=1}^{N} \left(\tau(X_i) - \tau\right)^2 + o_p(1).$$

Applying the law of large numbers gives $|\widehat{V}^{\tau(X)} - V^{\tau(X)}| = o_p(1)$, which completes the proof. □

B Weighted bootstrap derivation

To derive the form of the individual $\widetilde{\tau}_i$ for the weighted bootstrap, we start by writing $\widetilde{\tau}$ in terms of the naive matching estimator $\widehat{\tau}$ and the two bias-corrections $\widehat{B}_L^m$ and $\widehat{B}_L^a$.

$$\widetilde{\tau} = \widehat{\tau} - \widehat{B}_L^m - \widehat{B}_L^a \tag{31}$$

$$= \frac{1}{N}\sum_{i=1} \widetilde{\tau}_i \tag{32}$$

$$= \frac{1}{N}\sum_{i=1}^{N} \left(\widehat{\tau}_i - \widehat{B}_{Li}^m - \widehat{B}_{Li}^a\right) \tag{33}$$

$$= \frac{1}{N}\sum_{i=1}^{N} \widehat{\tau}_i - \frac{1}{N}\sum_{i=1}^{N} \widehat{B}_{Li}^m - \frac{1}{N}\sum_{i=1}^{N} \widehat{B}_{Li}^a \tag{34}$$

First, to derive $\widehat{\tau}_i$, we write the naive matching estimator as:

$$\hat{\tau} = \frac{1}{N}\sum_{i=1}^{N}(2A_i - 1)\left(1 + \frac{K_L^a(i)}{L}\right)\widehat{Y}_{i0} \tag{35}$$

$$= \frac{1}{N}\sum_{i=1}^{N}(2A_i - 1)\left[(1 - M_i)\left(1 + \frac{K_L^a(i)}{L}\right)Y_i + M_i\left(1 + \frac{K_L^a(i)}{L}\right)\frac{1}{L}\sum_{j\in\mathcal{J}_L^m(i)}Y_j\right] \tag{36}$$

$$= \frac{1}{N}\sum_{i=1}^{N}(2A_i - 1)\left[(1 - M_i)\left(1 + \frac{K_L^a(i)}{L} + \frac{K_L^m(i)}{L} + \frac{K_L^{am}(i)}{L^2}\right)Y_i\right] \tag{37}$$

$$\hat{\tau}_i = (2A_i - 1)\left[(1 - M_i)\left(1 + \frac{K_L^a(i)}{L} + \frac{K_L^m(i)}{L} + \frac{K_L^{am}(i)}{L^2}\right)Y_i\right] \tag{38}$$

with the second to last line following from the fact that all units used for matching imputation in the first stage have $M_i = 0$.

The first-stage bias correction, $\widehat{B}_L^m$ can also be written as

$$\widehat{B}_L^m = \frac{1}{N}\sum_{i=1}^{N}(2A_i - 1)\left(1 + \frac{K_L^a(i)}{L}\right)M_i\left(\frac{1}{L}\sum_{\ell\in\mathcal{J}_l^m(i)}\widehat{\mu}_{A_i0}(X_\ell, Z_\ell, A_i) - \widehat{\mu}_{A_i0}(X_i, Z_i, A_i)\right) \tag{39}$$

$$= \frac{1}{N}\sum_{i=1}^{N}(2A_i - 1)\left[\left(1 + \frac{K_L^a(i)}{L}\right)M_i\left(\frac{1}{L}\sum_{\ell\in\mathcal{J}_l^m(i)}\widehat{\mu}_{A_i0}(X_\ell, Z_\ell, A_i)\right) - M_i\left(1 + \frac{K_L^a(i)}{L}\right)\widehat{\mu}_{A_i0}(X_i, Z_i, A_i)\right] \tag{40}$$

$$= \frac{1}{N}\sum_{i=1}^{N}(2A_i - 1)\left[(1 - M_i)\left(\frac{K_L^m(i)}{L} + \frac{K_L^{am}(i)}{L^2}\right)\widehat{\mu}_{A_i0}(X_i, A_i, Z_i) - M_i\left(1 + \frac{K_L^a(i)}{L}\right)\widehat{\mu}_{A_i0}(X_i, A_i, Z_i)\right] \tag{41}$$

$$= \frac{1}{N}\sum_{i=1}^{N}(2A_i - 1)\widehat{\mu}_{A_i0}(X_i, A_i, Z_i)\left[(1 - M_i)\left(\frac{K_L^m(i)}{L} + \frac{K_L^{am}(i)}{L^2}\right) - M_i\left(1 + \frac{K_L^a(i)}{L}\right)\right] \tag{42}$$

$$\widehat{B}_{Li}^m = (2A_i - 1)\widehat{\mu}_{A_i0}(X_i, A_i, Z_i)\left[(1 - M_i)\left(\frac{K_L^m(i)}{L} + \frac{K_L^{am}(i)}{L^2}\right) - M_i\left(1 + \frac{K_L^a(i)}{L}\right)\right] \tag{43}$$

Note that because of exact matching in the first stage on A_i, $A_i = A_\ell$ for all $\ell \in \mathcal{J}_L^m(i)$.

And finally the second-stage bias correction $\widehat{B}_L^a$ can be rewritten as

$$\widehat{B}_L^a = \frac{1}{N}\sum_{i=1}^{N}(2A_i-1)\left[\frac{1}{L}\sum_{j\in J_L^a(i)}\widehat{\mu}_{1-A_i,0}(X_i,1-A_i)-\widehat{\mu}_{1-A_i,0}(X_j,1-A_i)\right] \tag{44}$$

$$= \frac{1}{N}\sum_{i=1}^{N}(2A_i-1)\left(\widehat{\mu}_{1-A_i,0}(X_i,1-A_i)+\frac{K_L^a(i)}{L}\widehat{\mu}_{A_i,0}(X_i,A_i)\right) \tag{45}$$

$$\widehat{B}_{Li}^a = (2A_i-1)\left(\widehat{\mu}_{1-A_i,0}(X_i,1-A_i)+\frac{K_L^a(i)}{L}\widehat{\mu}_{A_i,0}(X_i,A_i)\right) \tag{46}$$

Combining the three linearized terms yields

$$\begin{aligned}\widetilde{\tau}_i = (2A_i-1)\Bigg[&(1-M_i)\left(1+\frac{K_L^a(i)}{L}+\frac{K_L^m(i)}{L}+\frac{K_L^{am}(i)}{L^2}\right)Y_i \\ &-\left((1-M_i)\left(\frac{K_L^m(i)}{L}+\frac{K_L^{am}(i)}{L^2}\right)-M_i\left(1+\frac{K_L^a(i)}{L}\right)\right)\widehat{\mu}_{A_i,0}(X_i,A_i,Z_i) \\ &-\left(\widehat{\mu}_{1-A_i,0}(X_i,1-A_i)+\frac{K_L^a(i)}{L}\widehat{\mu}_{A_i,0}(X_i,A_i)\right)\Bigg]\end{aligned} \tag{47}$$

C Simulation results for inverse propensity of treatment weighting estimator

We omit from the simulation in the main text the bias and RMSE results of the inverse propensity of treatment weighting (IPTW) estimator compared to both sequential g-estimation, linear regression and telescope matching. This is because we observed extremely poor performance of IPTW relative to the other methods, particularly under the incorrectly specified models. The complete results are presented below in SM.6. In terms of root-mean square error, IPTW performs significantly worse than both sequential g-estimation and the naive linear regression estimator, with the performance worsening as the sample sizes increase.

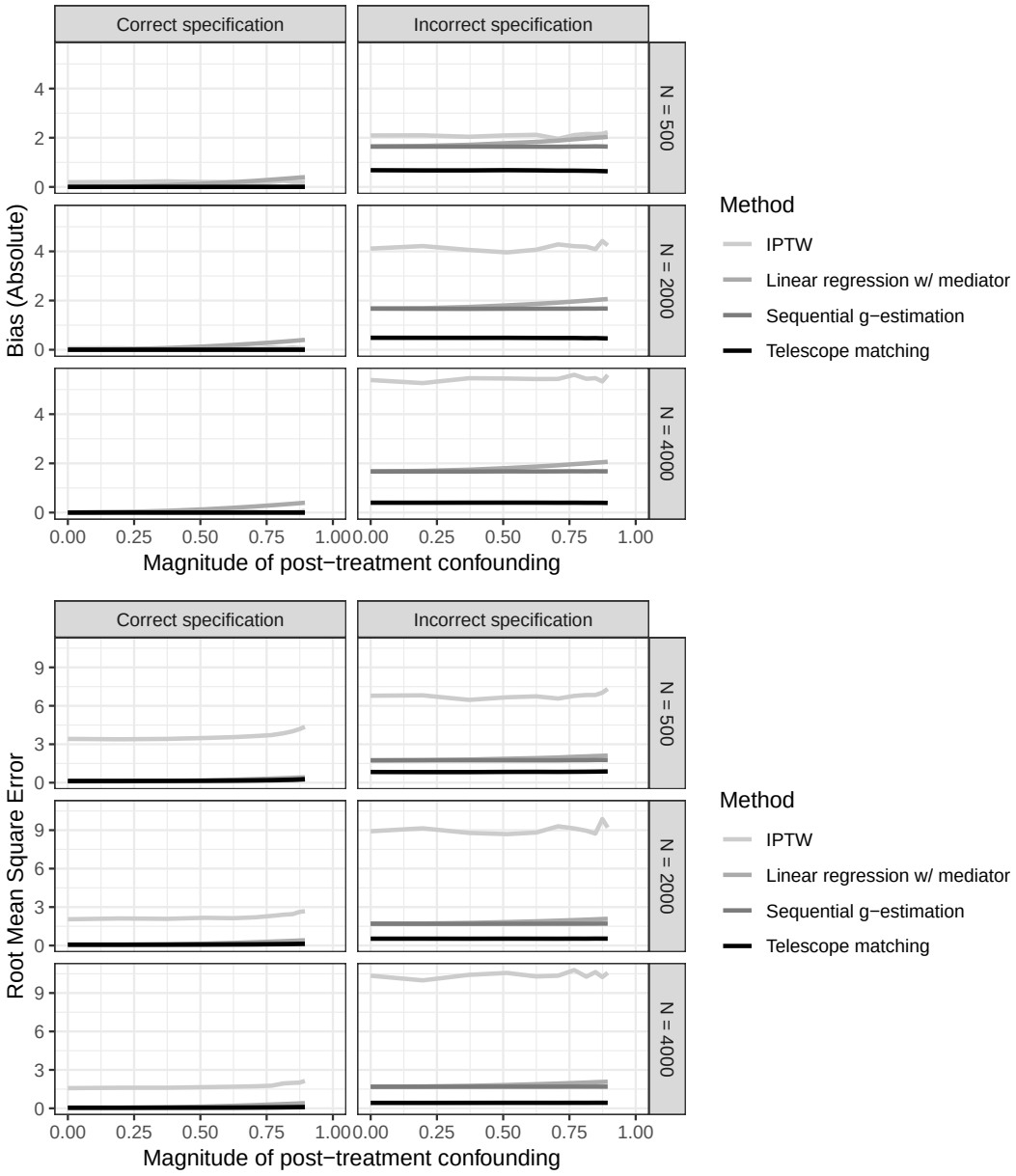

Figure SM.6: Performance of regression controlling for intermediate covariates, sequential g-estimation, inverse propensity of treatment weighting, and telescope matching under simulated data with correct and misspecified models

D Additional replication details

The replication data for Huber (2014) included 24 pre-treatment covariates and 16 post-treatment, pre-mediator covariates. Table SM.1 describes the label annotations that were provided for all of the variable names in the dataset. In the original analysis performed in the paper, the covariates all

entered into the propensity score models via a typical linear, additive specification.

Because treatment is randomly assigned, we are primarily concerned with imbalance on pre- and post-treatment covariates with respect to the mediator and the extent to which matching improves balance in the first stage. Table SM.2 reports the covariate means among men and women in each of the mediator groups (employed and unemployed between 1-1.5 years post-assignment). We notice the most stark imbalances arise in the pre-treatment employment indicators. Women who are employed 12-18 months after treatment are about 20 percentage points more likely to be employed in the year before the training program. Since past employment tends to predict future employment, this is certainly unsurprising. The imbalance is even more notable when we look at post-treatment measures of employment (Table SM.3). As should be expected, both women and men who are employed 12-18 months after treatment are about 50 percentage points more likely to be employed 9-12 months after treatment. This imbalance in employment history appears to be the most noticeable source of confounding of the mediator-outcome relationship.

We find that with a more parsimonious list of covariates related only to employment history, we can obtain nearly identical results for sequential g-estimation and telescope matching for women. Figure SM.7 plots the estimated ACDEs for women and men when we only adjust for two pre-treatment indicators: whether individuals were in school in the year before the program and whether they had a job in the year before the program and two post-treatment indicators: whether individuals worked 9-12 months after assignment at all and whether they had full time employment 9-12 after assignment. Notably, when we use a sequential-g estimator in which the mediator model is fully saturated, with all possible covariate interactions included, we obtain results substantively very close to those using telescope matching. This suggests that matching can improve robustness by lowering the consequences of model misspecification – when we allow for the most flexible parametric model, the differences between the two point estimates are negligible. In practice such fully saturated model specifications may be impossible (e.g. with continuous covariates) and suffer from problems of high variance, especially as the number of covariates grows. Therefore, the combination of non-parametric matching with parametric regression models may provide a reasonable compromise be-

tween efficiency and bias reduction.

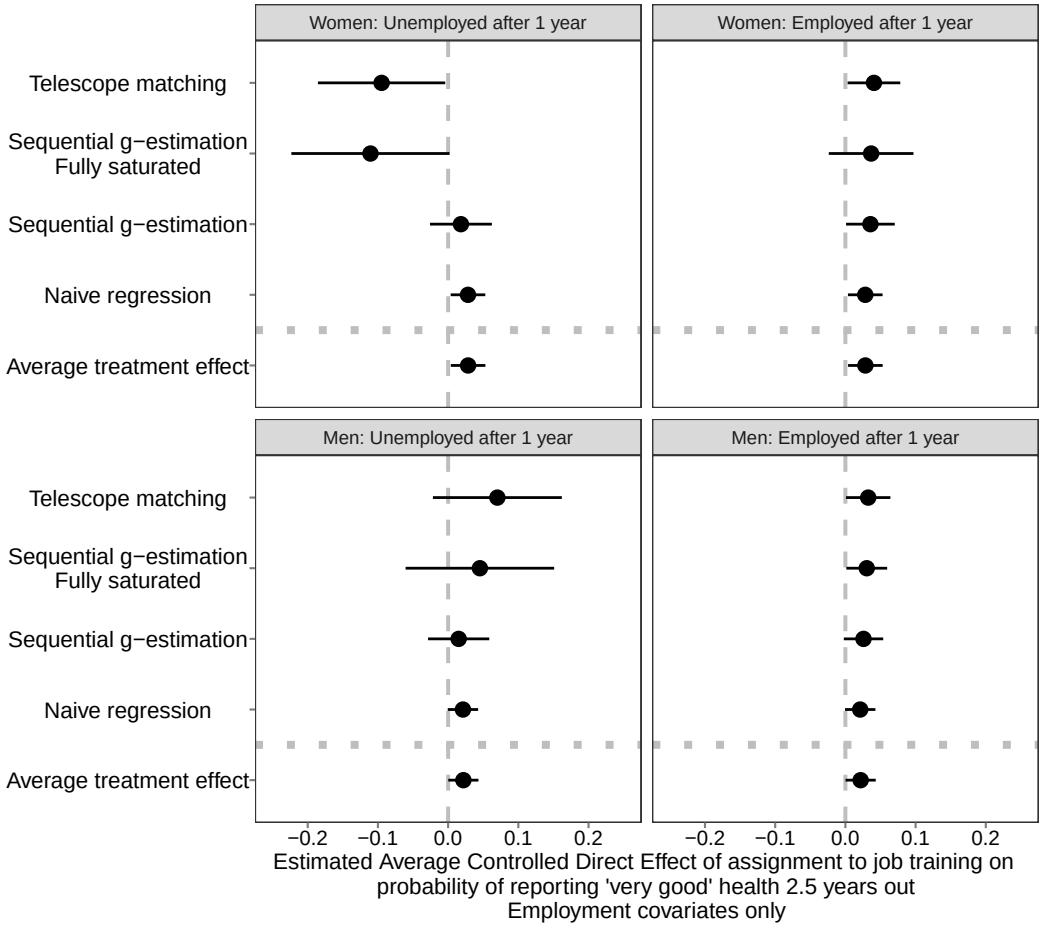

Figure SM.7: Estimated ACDEs of assignment to training among women (N = 4352) and men (N = 5673). Employment covariates only (2 pre-treatment, 2 post-treatment). Lines denote 90% confidence intervals. Sequential g-estimation CIs estimated using 1000 bootstrap iterations. Telescope matching CIs estimated using asymptotic variance estimator.

Pre-treatment	
Name	Label
schobef	in school 1yr before eligibility
trainyrbef	training in year before Job Corps
jobeverbef	ever had a job before Job Corps
jobyrbef	job in year before job corps
healtho12	good or very good health at assignment
healthomis	general health at assignment missing
pe_prbo	physical/emotional problems at assignment
pe_prbomis	missing - physical/emotional problems at assignment
everalc	ever abused alcohol before assignment
alc12	alcohol abuse one yr after assignment
everilldrugs	ever took illegal drugs before assignment
age_cat	age at application in years 16-24
edumis	education missing
eduhigh	higher education
rwhite	race - white
everarr	ever arrested before Job Corps
hhsize	household size at assignment
hhsizemis	missing - household size at assignment
hhinc12	low household income at assignment
hhinc8	high household income at assignment
fdstamp	received foodstamps in yr before assignment
welf1	once on welfare while growing up
welf2	twice on welfare while growing up
publicass	public assistance in yr before assignment
Post-treatment	
Name	Label
emplq4	worked some time 9-12 months after assignment
emplq4full	worked all the time in 9-12 months after assignment
pemplq4	proportion of weeks worked 9-12 months after assignment
pemplq4mis	missing - proportion of weeks worked 9-12 months after assignment
vocq4	in vocational training 9-12 months after assignment
vocq4mis	missing - in vocational training 9-12 months after assignment
health1212	very good or good health 1 yr after assignment
health123	fair health 1 yr after assignment
pe_prb12	1=phys/emot probs at 12 mths 0=no prob
pe_prb12mis	missing - physical/emotional problems 1 yr after assignment
narry1	number of arrests in year 1
numkidhhf1zero	no own kids living in household 1 yr after assignment
numkidhhf1onetwo	one or two own kids living in household 1 yr after assignment
pubhse12	1 = in public housing 1 yr after assignment, 0 = not in
h_ins12a	afdc and other transfers one yr after assignmen
h_ins12amis	missing - afdc and other transfers one yr after assignment

Table SM.1: Variable names and labels in Huber (2014)

	Women			Men		
Variable Name	Employed	Unemployed	Diff.	Employed	Unemployed	Diff.
schobef	0.644	0.696	-0.052	0.629	0.635	-0.006
trainyrbef	0.019	0.019	0.000	0.017	0.017	-0.000
jobeverbef	0.123	0.155	-0.032	0.132	0.183	-0.051
jobyrbef	0.728	0.526	0.202	0.707	0.480	0.227
healtho12	0.876	0.846	0.030	0.855	0.814	0.041
healthomis	0.015	0.028	-0.013	0.014	0.025	-0.011
pe_prbo	0.044	0.047	-0.004	0.054	0.055	-0.001
pe_prbomis	0.015	0.029	-0.015	0.014	0.024	-0.010
everalc	0.654	0.553	0.101	0.555	0.460	0.094
alc12	0.366	0.245	0.121	0.239	0.159	0.080
everilldrugs	0.005	0.006	-0.001	0.004	0.004	0.000
age_cat	18.506	17.880	0.627	18.750	18.377	0.373
edumis	0.014	0.026	-0.012	0.013	0.023	-0.010
eduhigh	0.024	0.011	0.013	0.040	0.020	0.020
rwhite	0.354	0.223	0.131	0.247	0.152	0.094
everarr	0.317	0.329	-0.012	0.160	0.172	-0.013
hhsize	4.271	4.359	-0.088	4.425	4.702	-0.276
hhsizemis	0.018	0.030	-0.011	0.017	0.026	-0.008
hhinc12	0.227	0.288	-0.062	0.313	0.382	-0.069
hhinc8	0.366	0.405	-0.040	0.329	0.393	-0.063
fdstamp	0.337	0.423	-0.086	0.506	0.593	-0.086
welf1	0.475	0.392	0.083	0.429	0.358	0.072
welf2	0.208	0.196	0.013	0.209	0.167	0.042
publicass	0.240	0.262	-0.022	0.251	0.283	-0.032

Table SM.2: Mean balance on pre-treatment covariates across mediator (employment after 12-18 mo.) by gender

	Women			Men		
Variable Name	Employed	Unemployed	Diff.	Employed	Unemployed	Diff.
emplq4	0.743	0.225	0.518	0.720	0.200	0.520
emplq4full	0.394	0.014	0.381	0.384	0.009	0.375
pemplq4	57.532	10.536	46.996	55.557	9.169	46.388
pemplq4mis	0.011	0.023	-0.012	0.009	0.017	-0.008
vocq4	0.186	0.231	-0.045	0.208	0.257	-0.048
vocq4mis	0.018	0.027	-0.009	0.013	0.015	-0.002
health1212	0.830	0.818	0.012	0.797	0.781	0.016
health123	0.123	0.127	-0.005	0.159	0.160	-0.000
pe_prb12	0.122	0.110	0.012	0.152	0.134	0.018
pe_prb12mis	0.035	0.038	-0.003	0.029	0.037	-0.008
narry1	0.240	0.335	-0.095	0.071	0.056	0.015
numkidhhf1zero	0.908	0.936	-0.028	0.670	0.576	0.094
numkidhhf1onetwo	0.087	0.060	0.027	0.300	0.376	-0.076
pubhse12	0.112	0.160	-0.048	0.157	0.219	-0.062
h_ins12a	1.160	1.225	-0.064	1.414	1.570	-0.156
h_ins12amis	0.091	0.117	-0.026	0.048	0.062	-0.014

Table SM.3: Mean balance on post-treatment covariates across mediator (employment after 12-18 mo.) by gender