Received: 11 March 2020 | Accepted: 31 August 2021

DOI: 10.1111/rssa.12759

ROYAL STATISTICAL SOCIETY
Series A Statistics in Society

ORIGINAL ARTICLE

Telescope matching for reducing model dependence in the estimation of the effects of time-varying treatments: An application to negative advertising

Matthew Blackwell[1] | **Anton Strezhnev**[2]

[1]Department of Government, Harvard University, Cambridge, Massachusetts, USA

[2]Department of Political Science, University of Chicago, Chicago, Illinois, USA

Correspondence
Matthew Blackwell, Department of Government, Harvard University, Cambridge, Massachusetts, USA.
Email: mblackwell@gov.harvard.edu

Abstract

Time-varying treatments are prevalent in the social sciences. For example, a political campaign might decide to air attack ads against an opponent, but this decision to go negative will impact polling and, thus, future campaign strategy. If an analyst naively applies methods for point exposures to estimate the effect of earlier treatments, this would lead to post-treatment bias. Several existing methods can adjust for this type of time-varying confounding, but they typically rely on strong modelling assumptions. In this paper, we propose a novel two-step matching procedure for estimating the effect of two-period treatments. This method, *telescope matching*, reduces model dependence without inducing post-treatment bias by using matching with replacement to impute missing counterfactual outcomes. It then employs flexible regression models to correct for bias induced by imperfect matches. We derive the asymptotic properties of the telescope matching estimator and provide a consistent estimator for its variance. We illustrate telescope matching by investigating the effect of negative campaigning in US Senate and gubernatorial elections. Using the method, we uncover a positive effect on turnout of negative ads early in a campaign and a negative effect of early negativity on vote shares.

KEYWORDS
causal inference, controlled direct effects, matching, time-varying treatments

1 | INTRODUCTION

Political campaigns are inherently dynamic. Candidates develop strategies in response to on-the-ground campaign conditions, which in turn affect strategies as races move closer to election day. A prominent example of this dynamic structure is the use of negative advertising. Candidates often 'go negative' in response to poor polling or to respond to attacks made by their opponents. But what effect does 'going negative' actually have? From the perspective of statisticians and analysts, estimating the time-varying effects of campaign negativity on various outcomes such as vote shares and voter turnout is challenging. Polling, for example, is both a consequence and cause of negativity at different points in time. And if polling is related to the outcome (voter turnout or vote shares in favour of a candidate), then controlling for polling earlier in a campaign cycle can lead to post-treatment bias in the estimation of the effect of early campaign negativity for a fixed level of late campaign negativity (Blackwell, 2013; Robins, 1997).

What can analysts do in these situations? Several parametric and semi-parametric methods have been developed for estimating such time-varying effects, including the parametric g-formula, structural nested models and marginal structural models (Richardson & Rotnitzky, 2014; Robins, 1986). Unfortunately, these extant approaches require the (correct) specification of several models, meaning that any inferences are heavily dependent on those modelling choices. Matching, on the other hand, is a popular strategy for estimating average treatment effects for a single binary treatment and is known to reduce model dependence (Abadie & Imbens, 2006; Dehejia & Wahba, 1999; Ho et al., 2006; Rosenbaum, 1995). However, although it has been widely adopted in the social and biomedical sciences due to its nonparametric and intuitive nature, its application has been very limited in situations with multiple or time-varying treatments. Standard matching can be used to estimate the effect of the last period treatment (since that can be formulated as a point-exposure problem), but it will fail for the effect of earlier exposure due to post-treatment bias.

In this paper, we present a new matching method for estimating the effect of time-varying treatments that helps reduce dependence on these modelling assumptions without inducing post-treatment bias. We focus on the case with two treatment exposures—in our example, negativity early in the campaign and late in the campaign. Our method matches in two steps, first for the early treatment (going negative early in a campaign) and then for the late treatment (going negative later), using different covariate sets for each step. This two-stage approach, which we call *telescope matching*, adjusts for only baseline confounders in the first stage, *telescoping out* to adjust for both the baseline and intermediate confounders in the second stage. These matching steps help impute missing counterfactual outcomes for each unit, which then can be used to estimate the effects of interest—in our case, the effect of a candidate going negative on voter turnout and the vote shares of the respective candidates.

Previous propensity score matching methods have been developed for time-varying treatments (Lechner, 2004), and we make two key contributions to this literature. First, we extend previous matching algorithms to handle direct matching on covariates rather than simply relying on propensity scores, which have been shown to have poor performance in matching applications (King & Nielsen, 2019). Second, we derive the large-sample properties of the telescope matching estimator, which can encompass both direct and propensity score matching, under a

fixed number of matches and show that while it is consistent for the effect of early exposure, it possesses a bias due to inexact matches that prevents convergence to a stable asymptotic distribution, as is the case with point-exposure matching (Abadie & Imbens, 2006). This bias even extends to propensity score matching with time-varying treatment, unlike in the single-treatment setting (Abadie & Imbens, 2016). 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 matching estimator.

Although 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 estimating these effects (such as structural nested mean models) more robust to modelling assumptions. We show that this is the case in our simulations—when the regression models are correctly specified, telescope matching shows similarly small levels of bias compared to these other methods but is slightly less efficient. When these models are misspecified, however, our procedure shows considerably lower bias even in a setting where other methods that attempt to reduce model dependence such as covariate balancing propensity scores are unstable (Imai & Ratkovic, 2015). While these results are not likely to hold across all possible data generating processes, they demonstrate that telescope matching can help to guard against misspecification of the outcome or propensity score models at the expense of efficiency when those models are correct.

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. We begin with a description of the applied setting of a candidate going negative in the course of a political campaign in Section 2. In Section 3, we define the relevant quantities of interest and the assumptions necessary to identify these effects—in our case, the effect of going negative on election-day outcomes. We then develop our telescope matching approach to estimating these direct effects, discuss its large-sample properties, describe the bias-correction approach, and derive variance estimators. In addition, we discuss how this matching estimator compares to other ways of estimating time-varying effects, with special attention to previously proposed sequential propensity score matching estimators. We then conduct a simulation study in Section 4 that shows how these various estimators perform when a researcher has correct and incorrect specifications of the outcome regression and propensity score models. In Section 5 we apply the method to our application of negative campaigning; using this technique, we find, contrary to the existing work in this area, that there is a positive effect of an incumbent candidate going negative early in the campaign on voter turnout. We also find a negative effect on early negativity on incumbent vote *shares*. This suggests that 'going negative' might actually create a backlash effect where outpartisans are motivated to turnout against a candidate who runs the negative ads.

2 | MOTIVATING APPLICATION

A substantial literature in political science addresses the question of 'campaign effects'—that is, how the course of political campaigns affect various electoral outcomes (see Jacobson, 2015, for a review of this literature). Here, scholars have studied a wide variety of tools that campaigns

use to mobilize (or demobilize) voters and to persuade voters to change their vote. These tools include door-to-door canvassing, rallies, speeches, news conferences, social media posts and advertisements.

One particularly important campaign tool is negative campaigning—or, ads that directly attack an opponent. (This stands in contrast to ads that simply promote the candidate herself.) Although a number of studies using experimental and observational approaches have tried to estimate the effect of negative advertising on turnout in US elections (Lau et al., 2007), the direction of the effect is difficult to predict *a priori*. On the one hand, we might think that negativity generates more media attention, leading to more interest in the campaign and thus higher turnout. On the other hand, negativity might cause citizens to become disenchanted with the political process and, thus, disengage, leading to lower voter turnout. Evidence from the empirical literature is mixed. One meta-analysis concluded that 'the research literature provides no general support for the hypothesis that negative political campaigning depresses voter turnout. If anything, negative campaigning more frequently appears to have a slight mobilizing effect' (Lau et al., 2007).

A problem with these studies is, however, that they fail to consider the dynamic nature of campaigning. Candidates tend to 'go on the attack' in response to something—usually something like falling behind in the polls or being on the receiving end of attack ads from their own opponents (Blackwell, 2013). Ignoring the dynamic nature of negative advertising means that existing studies tend to estimate the effect of 'going negative' at just one point in time. Not only does this approach ignore how effects vary over time, but also it becomes challenging to think about which variables to control for and which to ignore.

Here, we investigate the effect of campaign negativity on voter turnout and vote shares in US Senate and gubernatorial elections from 2000 to 2016. For each campaign, we have collected data on the types of television ads shown, the polling of the candidates in the race, the amount donated to each candidate in the race, and a host of background features of the race and its candidates. Because we are interested in the effects of negativity early in the race (from the primary until the end of September), we need a method that can adjust for time-varying confounding without introducing post-treatment bias. Furthermore, many of the covariates we measure are continuous with potentially nonlinear relationships with both the decision to go negative and the outcomes. Thus, it is essential to have a method that is robust to the sometimes ad-hoc modelling decisions researchers are forced to make.

3 | PROPOSED METHOD

3.1 | Notation and assumptions

We focus on the case of two binary treatment exposures over time, though it is possible to extend the approach to arbitrary numbers of periods. Let $A_{i1} \in \{0, 1\}$ and $A_{i2} \in \{0, 1\}$ denote the value of early and late treatment, respectively, for unit i. The goal of the analysis is to estimate the effect of treatment on some outcome, Y_i. We define potential outcomes for this variable under the various combinations of the treatment history, $Y_i(a_1, a_2)$ (Robins, 1986; Rubin, 1974). We make the usual consistency assumption, $Y_i = Y_i(a_1, a_2)$ if $A_{i1} = a_1$ and $A_{i2} = a_2$, which states that the observed outcome for unit i is the potential outcome for that unit at its observed level of A_{i1} and A_{i2}.

We define two sets of relevant covariates: baseline and intermediate. The baseline covariates, X_i, are causally prior to both treatments. 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_{i1}, but are causally prior to A_{i2} and confound the relationship between the outcome and late treatment. These covariates pose problems for standard models when trying to estimate the effect of the entire treatment history due to the potential for post-treatment bias induced by conditioning on them (Robins, 1997; Rosenbaum, 1984). An example of this causal structure is shown in Figure 1.

Our goal in this paper is to estimate the effect of early exposure to negative ads on voter turnout for a fixed value of late negativity: $\tau_{a_2} = E\{Y_i(1, a_2) - Y_i(0, a_2)\}$. This quantity represents the average effect of early treatment when late treatment is fixed at a particular value. For this paper, we focus on the effect when $a_2 = 0$ and write $\tau \equiv \tau_0$, but it is straightforward to estimate the effects at any level. We can also define the conditional effect of early treatment: $\tau(x) = E\{Y_i(1, 0) - Y_i(0, 0) \mid X_i = x\}$. This is the direct effect of early treatment within levels of the baseline covariates. We can recover the marginal effect from the conditional effects by averaging over the distribution of the data: $\tau = E\{\tau(X_i)\}$. In the context of mediation studies, this quantity is also known as the controlled direct effect (Robins & Greenland, 1992) (see the dashed lines in Figure 1).

We make the following sequential ignorability assumption about the treatment history:

Assumption 1 (Sequential Ignorability) For every value, a_1, a_2, x, z:

$$\{Y_i(a_1, a_2), Z_i(a_1)\} \perp\!\!\!\perp A_{i1} | X_i = x, \tag{1}$$

$$Y_i(a_1, a_2) \perp\!\!\!\perp A_{i2} | X_i = x,\ Z_i = z,\ A_{i1} = a_1. \tag{2}$$

The first part of this assumption states that early treatment is independent of the potential outcomes, conditional on baseline covariates. The second part states that the late treatment is independent of the potential outcomes, conditional on early treatment and the baseline and intermediate covariates. This assumption essentially requires two 'selection-on-observables' conditions, one for each treatment. Thus, there must be no unmeasured confounders for the early-treatment-outcome relationship after conditioning on X_i and no unmeasured confounders for the late-treatment-outcome relationship after conditioning on $\{X_i, A_{i1}, Z_i\}$. Note that this sequential ignorability assumption is considerably weaker than a version of sequential ignorability used in mediation analyses that requires no intermediate confounders (Imai et al., 2010).

We further assume that the distributions of the treatments are not degenerate at any values of the covariates.

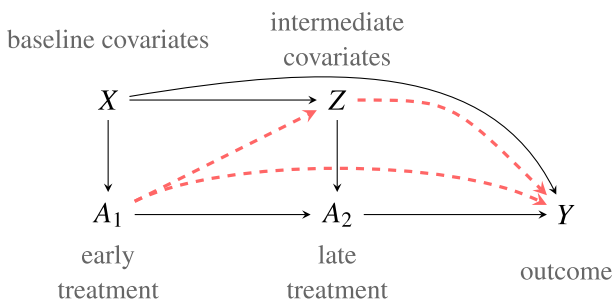

FIGURE 1 Directed acyclic graph showing the causal relationships in the present setting. Dashed red lines represent the effect of early treatment for fixed values of later treatment. Unobserved errors are omitted

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

$$\eta < \text{pr}(A_{i1} = 1 | X_i = x) < 1 - \eta, \tag{3}$$

$$\nu < \text{pr}(A_{i2} = 1 | X_i = x,\ Z_i = z,\ A_{i1} = a) < 1 - \nu. \tag{4}$$

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 $A_{i2} = 1$ and $A_{i2} = 0$ covariate distributions. These are straightforward generalizations of the common support assumptions in the matching literature to this 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 $\mu_{a_1 a_2}(x, z, a_1) = E\{Y(a_1, a_2) | X_i = x, Z_i = z, A_{i1} = a_1\}$ and $\mu_{a_1 a_2}(x, a_1) = E\{Y_i(a_1, a_2) | X_i = x, A_{i1} = a_1\}$. Let $\mu(x, z, a_1, a_2) = E(Y_i | X_i = x, Z_i = z, A_{i1} = a_1, A_{i2} = a_2)$ be the CEF of the observed outcome, noting that under Assumption 1, $\mu_{a_1 a_2}(x, z, a_1) = \mu(x, z, a_1, a_2)$. We also define two types of residuals, $\varepsilon_i = Y_i - \mu(X_i, Z_i, A_{i1}, A_{i2})$ and $\eta_i = \mu_{A_{i1}0}(X_i, Z_i, A_{i1}) - \mu_{A_{i1}0}(X_i, A_{i1})$. 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 $E(\eta_i | \mathbf{X}, \mathbf{A}_1) = 0$ and $E(\varepsilon_i | \mathbf{X}, \mathbf{Z}, \mathbf{A}_1, \mathbf{A}_2) = 0$, where $\mathbf{X}$ and $\mathbf{Z}$ are the entire $N \times k_x$ and $N \times k_z$ matrices of baseline and intermediate covariates, and $\mathbf{A}_1$ and $\mathbf{A}_2$ are the N vectors of the early and late treatments. Finally, we define various conditional variance functions. Let $\sigma^2(x, z, a_1, a_2) = \text{var}(Y_i | X_i = x, Z_i = z, A_{i1} = a_1, A_{i2} = a_2)$ and $\sigma^2_\eta(x, a_1) = \text{var}[E\{Y_i(a_1, 0) | X_i = x, Z_i, A_{i1} = a_1\} | X_i = x, A_{i1} = a_1] = E(\eta_i^2 | X_i = x, A_{i1} = a_1)$. Again, under Assumption 1, $\sigma^2(x, z, a_1, a_2) = \text{var}\{Y_i(a_1, a_2) | X_i = x, Z_i = z, A_{i1} = a_1\}$.

3.2 | The telescope matching procedure

How can we estimate the effect of early treatment fixing late treatment? If A_{i1} and A_{i2} are jointly randomized, then standard tools for multileveled treatments can be used to estimate the direct effect of a treatment since there are no covariates for which to adjust. When there are only baseline confounders, then standard selection-on-observable methods for multileveled treatments can be applied (Imbens, 2004). However, when there are post-treatment confounders for the relationship between A_{i2} 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 effect of A_{i1} when $A_{i2} = 0$, 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. We describe the technical details of the matching and imputation procedure here, but we also provide a simple example of the procedure with $N = 6$ in Supplemental Materials Section A.

The broad approach to sequential matching for time-varying treatments we describe below was first proposed by Lechner (2004) and focused on creating pair matches using the propensity scores for A_{i1} and A_{i2} (see Huber et al., 2018; Lechner & Miquel, 2010, for applications of this method). Our proposed matching approach generalizes this algorithm to allow for using more than a single match and to allow for matching directly on covariates in addition to propensity

scores. Generalizing beyond a single match is important for trading off bias and efficiency of our estimators, while allowing direct matching provides robustness against misspecification of the propensity score models. Below we discuss more fully the advantages and disadvantages of working with direct matching versus propensity score matching. Finally, we extend the proposal of Lechner (2004) by deriving the large-sample properties of telescope matching in the following sections and show that all of these matching methods require bias correction.

In the first step of telescope matching, we match each unit to L units of the opposite early treatment status with similar values of the baseline covariates as if we were attempting to estimate the average treatment effect of A_{i1} adjusting for X_i. We follow Abadie and Imbens (2006) in much of our discussion of matching estimators. Given a particular distance metric on the support of X_i (such as the Euclidean norm or the Mahalanobis distance) and given a particular unit i, we choose L units, here indexed by j, that are the closest to i in terms of covariate distance that have $A_{j1} = 1 - A_{i1}$. Let $\mathcal{J}_{1L}(i)$ denote this set of units that are matched to some unit i. Matching is done with replacement so each unit might be matched to multiple times, and we let $K_{1L}(i) = \sum_{k=1}^{N} \mathbb{I}\{i \in \mathcal{J}_{1L}(k)\}$ be the number of times that unit i is used as a match in the first 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.

In the second stage, we match across levels of late treatment to minimize the imbalance in terms of both baseline and intermediate confounders. Letting $V_i = (X_i, Z_i)$, the second step of telescope matching is to match each unit with $A_{i2} = 1$ to some number of units, indexed by ℓ, with $A_{\ell 2} = 0$ that have similar values of covariates V_i and identical early treatment status $A_{i1} = A_{\ell 1}$. Let $\mathcal{J}_{2L}(i)$ denote this set of units that are matched to some unit i with $A_{i2} = 1$, with $K_{2L}(i) = \sum_{k=1}^{N} \mathbb{I}\{i \in \mathcal{J}_{2L}(k)\}$ being the number of times that unit i is used as a match in the second stage.

Typically, each of these steps would be used to estimate the average effect of treatment given the past, but here we are actually more interested in obtaining a good imputation of the potential outcome for different treatment histories. We do so by moving backward through the treatment history. Let $Y_i(A_{i1}, 0)$ be the composite potential outcome under the observed early treatment status for i, but with late treatment set to 0, which is unobserved for any unit with $A_{i2} = 1$. We define the following imputation:

$$\widehat{Y}_i(A_{i1}, 0) = \begin{cases} Y_i & \text{if } A_{i2} = 0 \\ \dfrac{1}{L} \displaystyle\sum_{\ell \in \mathcal{J}_{2L}(i)} Y_\ell & \text{if } A_{i2} = 1 \end{cases}$$

For units observed with $A_{i2} = 0$, we observe $Y_i(A_{i1}, 0) = Y_i$ by consistency. However, for units with $A_{i2} = 1$, we need to impute the missing counterfactual outcome and do so by averaging the outcome among those units with $A_{i2} = 0$ which were matched to unit i. These units have identical early treatment levels A_{i1} and are the closest to i in terms of the baseline and intermediate covariates. Using the first-stage matching, we can now generate imputations of the base potential outcomes with $A_{i2} = 0$:

$$\widehat{Y}_i(a_1, 0) = \begin{cases} \widehat{Y}_i(A_{i1}, 0) & \text{if } A_{i1} = a_1 \\ \dfrac{1}{L} \displaystyle\sum_{j \in \mathcal{J}_{1L}(i)} \widehat{Y}_j(A_{j1}, 0) & \text{if } A_{i1} = 1 - a_1 \end{cases}$$

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 effect of early treatment then becomes

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

In general, this estimator will perform well when the matching discrepancies in the covariates are small. One can assess the matching quality of each step using standard matching diagnostics applied to the appropriate covariate set. For instance, for the second step, we would check the balance of the matched $A_{i2} = 1$ units to the $A_{i2} = 0$ units in terms of both X_i and Z_i.

Both $K_{1L}(i)$ and $K_{2L}(i)$ 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 $A_{i2} = 0$ might also contribute *indirectly* if they are matched to a $A_{i2} = 1$ unit in the second stage and that $A_{i2} = 1$ unit is used as a match in the first stage. To account for such indirect contributions of a unit, let $K_{*L}(i) = \sum_{j=1}^{N} \mathbb{I}\{i \in \mathcal{J}_{2L}(j)\} K_{1L}(j)$ be the number of times a second-stage match with $A_{i2} = 0$ is implicitly used as a match in the first stage because the unit to which it was matched is selected as a match in the first stage. We can rewrite the simple telescope matching estimator as a weighted average of units with $A_{i2} = 0$, $\widehat{\tau} = N^{-1} \sum_{i=1}^{N} (2A_{i1} - 1)(1 - A_{i2}) W_i Y_i$, where:

$$W_i = 1 + \frac{K_{1L}(i)}{L} + \frac{K_{2L}(i)}{L} + \frac{K_{*L}(i)}{L^2}. \tag{5}$$

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. One data-driven approach to selecting L would be to start at some modest number such as $L = 5$ or $L = 10$ and then decrease the matching ratio until some pre-specified balance criteria is met. Unfortunately, given the variety of possible balance criteria, it is beyond the scope of this paper to develop a general way to find the optimal matching ratio.

3.3 | Bias and consistency

Abadie and Imbens (2006) showed that in the context of estimating the overall ATE of a point exposure, the equivalent simple matching procedure was biased due to imperfect matches. Furthermore, they showed that with a fixed size for the matched set, L, this bias converges to 0 as the sample size increases, but at a rate slow enough to affect the asymptotic distribution 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:

$$\widehat{\tau} - \tau = \left(\frac{1}{N} \sum_{i=1}^{N} \tau(X_i) - \tau \right) + U_{1L} + U_{2L} + B_{1L} + B_{2L}. \tag{6}$$

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 effects and the marginal effect, which converges to 0 under a standard law of large numbers. Next in the decomposition are two weighted sums of the residuals:

$$U_{2L}=\frac{1}{N}\sum_{i=1}^{N}(2A_{i1}-1)(1-A_{i2})\left(1+\frac{K_{1L}(i)}{L}+\frac{K_{2L}(i)}{L}+\frac{K_{*L}(i)}{L^2}\right)\varepsilon_i,$$
$$U_{1L}=\frac{1}{N}\sum_{i=1}^{N}(2A_{i1}-1)\left(1+\frac{K_{1L}(i)}{L}\right)\eta_i.$$

The error due to the second-stage is mean-zero conditional on all variables, $E(U_{2L}\,|\,\mathbf{X},\,\mathbf{A}_1,\,\mathbf{Z},\,\mathbf{A}_2)=0$, and the error due to the first-stage is mean-zero conditional on the baseline covariates and the treatment, $E(U_{1L}\,|\,\mathbf{X},\,\mathbf{A}_1)=0$. 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:

$$B_{2L}=\frac{1}{N}\sum_{i=1}^{N}(2A_{i1}-1)A_{i2}\left(1+\frac{K_{1L}(i)}{L}\right)\left(\frac{1}{L}\sum_{\ell\in J_{2L}(i)}\mu_{A_{i1}0}(X_\ell,Z_\ell,A_{i1})-\mu_{A_{i1}0}(X_i,Z_i,A_{i1})\right),$$
$$B_{1L}=\frac{1}{N}\sum_{i=1}^{N}(2A_{i1}-1)\left[\frac{1}{L}\sum_{j\in J_{1L}(i)}\mu_{1-A_{i1},0}(X_i,1-A_{i1})-\mu_{1-A_{i1},0}(X_j,1-A_{i1})\right].$$

These bias terms reflect the matching discrepancy at each stage of matching. For instance, the last term in the definition of B_{2L} 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 $A_{i2}=1$ unit is matched in the first stage. If the matches were perfect, then we would have $X_i=X_\ell$ and $Z_i=Z_\ell$ for all $\ell\in\mathcal{J}_{2L}(i)$ and $X_i=X_j$ for all $j\in\mathcal{J}_{1L}(i)$, 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 & 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:

1. 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.
2. $\{(Y_i,A_{i2},Z_i,A_{i1},X_i)\}_{i=1}^{N}$ are independent and identically distributed.
3. The functions $\mu(x,z,a_1,a_2)$, $\sigma^2(x,z,a_1,a_2)$, and $\sigma_\eta^2(x,a_1)$ are Lipschtiz on $\mathbb{V}$.
4. $E(Y_i^4\,|\,V_i=v,A_{i1}=a_1,A_{i2}=a_2)$ exists and is uniformly bounded in $\mathbb{V}$ for all a_1 and a_2.
5. $\sigma^2(x,z,a_1,a_2)$ and $\sigma_\eta^2(x,a_1)$ are bounded away from 0.

These assumptions 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 effect of early treatment.

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

$$
\begin{aligned}
V^{\tau(X)} &= E\left\{(\tau(X_i) - \tau)^2\right\}, \qquad V^{\eta} = E\left\{\left(1 + \frac{K_{1L}(i)}{L}\right)^2 \sigma_{\eta}^2(X_i, A_{i1})\right\}, \\
V^{\epsilon} &= E\left\{(1 - A_{i2})\left(1 + \frac{K_{1L}(i)}{L} + \frac{K_{2L}(i)}{L} + \frac{K_{*L}(i)}{L^2}\right)^2 \sigma^2(X_i, Z_i, A_{i1}, 0)\right\}.
\end{aligned} \tag{7}
$$

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 0 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 & 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 & Imbens, 2011).

There are several tuning parameters for matching that deserve attention in our setting. It is important to note that our approach does not prune observations for which the matching discrepancy on the covariates is large as does radius or kernel matching. While it is possible to prune based on baseline covariates, any pruning based on intermediate confounders has the potential to induce post-treatment bias and so we avoid this. For the sake of efficiency, it is possible to extend the theoretical results here to allow for a variable number of matches for each unit, so long as the minimum number of matches is one and therefore no pruning occurs (Abadie & Imbens, 2012). Of course, allowing for a large number of matches for a particular unit could increase the number of indirect matches, $K_{*L}(i)$, make the implied matching weights more variable and actually reduce efficiency. The additional benefit of a variable matching ratio in this setting would be a fruitful avenue for future research.

3.4 | Bias correction

Due to the large-sample bias of a simple matching estimator, Abadie and Imbens (2011) proposed a bias-corrected estimator that estimates and removes this bias. 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 conditional expectations, $\widehat{\mu}(x, z, a_1, a_2)$ and $\widehat{\mu}_{a_1 0}(x, a_1)$. As in Abadie and Imbens (2011), we leverage a flexible series estimator that grows more complex with the sample size, which we describe in more technical detail in the Supplemental Material Section B. We can then use these regressions to obtain estimates of the bias terms themselves:

$$\widehat{B}_{2L} = \frac{1}{N}\sum_{i=1}^{N}(2A_{i1}-1)\left(1+\frac{K_{1L}(i)}{L}\right)A_{i2}\left(\frac{1}{L}\sum_{\ell\in J_{2L}(i)}\widehat{\mu}(X_\ell, Z_\ell, A_{i1}, 0)-\widehat{\mu}(X_i, Z_i, A_{i1}, 0)\right),$$
$$\widehat{B}_{1L} = \frac{1}{N}\sum_{i=1}^{N}(2A_{i1}-1)\left[\frac{1}{L}\sum_{j\in J_{1L}(i)}\widehat{\mu}_{1-A_{i1},0}(X_i, 1-A_{i1})-\widehat{\mu}_{1-A_{i1},0}(X_j, 1-A_{i1})\right].$$

If the regression estimators are consistent for their respective conditional expectations (as they are under our regularity conditions), then $\widehat{B}_{2L}$ and $\widehat{B}_{1L}$ converge in probability to the bias terms B_{2L} and B_{1L} respectively. With these estimates in hand, we define the following bias-corrected telescope matching estimator: $\widetilde{\tau} \equiv \widehat{\tau} - \widehat{B}_{2L} - \widehat{B}_{1L}$. We establish the asymptotic distribution of this estimator in Theorem 2.

Theorem 2 (Bias-corrected matching) *Suppose that Assumptions 1–3 hold along with Assumption 1 in the Supplemental Material. Then,* $\sqrt{N}(B_{2L} + B_{1L} - (\widehat{B}_{2L} + \widehat{B}_{1L}))$ *converges in probability to 0 and* $\sqrt{N}(\widetilde{\tau} - \tau)$ *converges in distribution to* $N(0, \sigma^2)$.

Theorem 2 shows that, under some 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 there are relatively many $A_{i2} = 1$ units, since, intuitively, there is more bias to correct in that setting. In these situations σ^2 might not be a good approximation to the finite sample variance of $\widetilde{\tau}$.

3.5 | 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 Equation (7). One approach to estimating the variance of the matching estimator is to directly estimate the components of Equation (7), which are the variance of $\tau(X_i)$ and weighted averages of the conditional variances of the outcomes. A straightforward way to implement such an estimator is to replace the population 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:

$$\begin{aligned}
\widehat{V}^{\tau(X)} &= \frac{1}{N}\sum_{i=1}^{N}\left(\widehat{\mu}_{10}(X_i)-\widehat{\mu}_{00}(X_i)-\widetilde{\tau}\right)^2,\\
\widehat{V}^{\eta} &= \frac{1}{N}\sum_{i=1}^{N}\left(1+\frac{K_{1L}(i)}{L}\right)^2\left(\widehat{\mu}(X_i, Z_i, A_{i1}, 0)-\widehat{\mu}_{A_{i1}0}(X_i, A_{i1})\right)^2, \qquad (8)\\
\widehat{V}^{\varepsilon} &= \frac{1}{N}\sum_{i=1}^{N}\left(1-A_{i2}\right)W_i^2\left(Y_i-\widehat{\mu}(X_i, Z_i, A_{i1}, 0)\right)^2.
\end{aligned}$$

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 an estimator similar to the one 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) *Under the conditions of Theorem 2,* $\widehat{\sigma}^2 \xrightarrow{p} \sigma^2$.

While the bootstrap is a popular approach for many methods, it is well known that conventional non-parametric bootstrapping, resampling observations $\{Y_i, X_i, Z_i, A_{i1}, A_{i2}\}$, is invalid for matching estimators (Abadie & Imbens, 2008). This is due to the inability of the naive boostrap to preserve the distributions of $K_{1L}(i)$, 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: $K_{2L}(i)$ and $K_{*L}(i)$. 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. As discussed by Otsu and Rai (2017), this approach avoids the issues with the naive row-resampling bootstrap method that is commonly used and analysed by Abadie and Imbens (2008) in the matching context. We extend their procedure to our setting in the Supplemental Materials.

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 would 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 and numbers of covariates in Supplemental Material Section C. Overall, we find that for reasonable sample sizes and small numbers of covariates, our asymptotic variance estimator and the weighted bootstrapping provide a reliable method for constructing confidence intervals when using telescope matching. In contexts with more covariates, however, coverage drops for both methods, though the effect is much more pronounced for the weighted bootstrap. Coverage for our asymptotic variance estimator is very similar to the variance estimators of single-treatment, bias-corrected matching estimators of Abadie and Imbens (2006), suggesting this drop is broader feature of matching estimators rather than sometime specific to telescope matching.

3.6 | Matching on the propensity scores

When there are a large number of covariates, direct matching in the manner described here can be difficult. A widely used alternative to direct matching is matching on the estimated propensity score, or the probability of treatment given the covariates. This has the advantages of

being a simpler univariate matching problem while ensuring balance on the covariates when the propensity score model is correctly specified (Rosenbaum & Rubin, 1983). Abadie and Imbens (2016) studied the asymptotic properties of propensity score matching and derived asymptotic variance estimators under both known and estimated propensity scores.

Versions of the propensity score matching approach have been applied to the time-varying treatment setting. Lechner (2004) first proposed a sequential matching algorithm based on the propensity scores for A_{i1} and A_{i2}, but did not derive the formal properties of that estimator. More recently, Huber et al. (2018) use a matching approach to estimate controlled direct effects in a setting where there are no intermediate confounders. These sequential propensity score approaches are attractive because they reduce the dimensionality of the matching problem considerably, but they have four drawbacks in this setting. First, the Lechner (2004) algorithm is limited to one-to-one matching, which could lead to a loss in efficiency relative to matching algorithms that allow for multiple matches.

Second, the sequential matching of Lechner (2004) prunes observations based on propensity scores for A_{i2}, which can induce post-treatment bias since that is implicitly conditioning the analysis on a post-treatment variable Z_i. In particular, let $\pi_1(x) = \text{pr}(A_{i1} = 1 \,|\, X_i = x)$ be the propensity score function for $t = 1$ and $\pi_i(x, z, a) = \text{pr}(A_{i2} = 1 \,|\, X_i = x, Z_i = z, A_{i1} = a)$ be the propensity score function for $t = 2$. Then, this sequential matching approach will restrict estimation of the direct effects to units in a particular range of $\pi_{i2} = \pi_i(X_i, Z_i, A_{i1})$, which implicitly conditions on Z_i. While conditioning on these intermediate covariates is valid within an adjustment step (matching in our case, or conditional expectations for structural nested mean models), the final estimation of the direct effect should only be conditional on baseline covariates (Robins, 1997).

Third, as we discuss below, these previous methods did not incorporate bias correction for the asymptotic bias that comes from imperfect matches. As we show in the simulations, this bias can be substantial when the propensity score model is incorrectly specified. Direct matching, on the other hand, can still recover decent matches because the validity and quality of the match depends less on functional form assumptions. Finally, recent work has shown that the dimension reduction inherent in propensity score matching can actually increase imbalance and bias compared to direct matching methods (King & Nielsen, 2019). This last drawback urges caution and careful checking when deciding between propensity score and direct matching methods.

If propensity score matching is a better fit for a particular application, there is a straightforward way to incorporate propensity scores into telescope matching: simply include them as the sole baseline and intermediate covariates in the above method. The propensity score telescope matching procedure would take $\pi_{i1} = \pi_i(X_i)$ as the sole baseline covariate and $\pi_{i2} = \pi_i(X_i, Z_i, A_{i1})$ to be the sole intermediate covariate. With $L = 1$ and no bias correction, this procedure is essentially what is proposed by Lechner (2004). If the propensity score functions were known, this approach is valid since the balancing property of the propensity score ensures Assumption 1 holds. With mild restrictions on the propensity score, such as those in Abadie and Imbens (2016), the propensity scores will also satisfy Assumptions 2 and 3. Finally, we note that, because we have two propensity scores in matching for A_{i2}, there will be asymptotic bias in the estimator as described above and bias correction will be required as in the direct matching case. Theorem 1 of Abadie and Imbens (2006) showed that the conditional bias term will be $B_{1L} = O_p(N^{-1/2})$ and it will cause bias unless the conditional mean of the potential outcomes do not vary with the propensity scores, a very unrealistic assumption.

Accounting for the estimation of the propensity scores in the variance of the telescope matching procedure is much more difficult. Abadie and Imbens (2016) are able to derive these in the single-treatment setting, but unfortunately, those results of are not applicable with two treatments

because they rely on propensity score matching being univariate. This allows a simpler statement of the variance of the matching estimator in terms of the true propensity score and the variance when the propensity score is estimated. The telescoping nature of the confounding and the non-collapsibility of the propensity score means that the matching for A_{i2} must be matched on the propensity scores for both A_{i2} *and* A_{i2}, ensuring that the resulting estimator will be asymptotically biased. Because of this, it is not clear how to derive the variance of the telescope matching estimator when the propensity scores are estimated and bias correction is employed. However, Abadie and Imbens (2016) show that, when targeting average rather than conditional effects, the standard variance formulas will tend to be conservative—that is, they will overestimate the variance. A similar argument can be used to show that the propensity score-based telescope matching will have the same property since the average effects of interest here do not depend on the propensity score as with the ATE in Abadie and Imbens (2016). Alternatively, one could use the nonparametric bootstrap for both the estimation of the propensity score and matching analysis similar to how one could use it to account for bias correction above. Given the known theoretical shortcomings of the bootstrap, its performance will depend on the particular empirical setting.

3.7 | Relationship to other approaches

Time-varying treatments 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 A_{i2} and Z_i due to the potential post-treatment bias, which is sometimes called collider bias. The estimation of these effects has focused on two general approaches. First, structural nested mean models estimate τ by first modelling the (conditional) effect of A_{i2} on Y_i, and then removing that effect from the outcome (Robins, 1997). This 'blipped-down' outcome is similar to our imputed potential outcomes above, but this approach requires the correct parametric specification of $\mu(x, z, a_1, a_2)$ to consistently estimate τ. Telescope matching leverages estimates of this same conditional expectation, but the nonparametric matching feature of the approach should make it less sensitive to minor misspecification. Our simulations below bear out this conjecture. Structural nested mean models have two other limitations that make it difficult to use in all settings. First, if the effect of A_{i2} varies by Z_i, the approach requires the integration of the regression functions over the distribution of Z_i. This integration, though, requires parametric models for the outcome and for the joint distribution of the covariates, which can be very demanding when there are more than a handful of covariates. The telescoping nature of our matching approach sidesteps this integration. Second, structural nested mean models are difficult to apply to binary outcomes (Robins, 2000; Robins & Rotnitzky, 2004; Vansteelandt, 2010), whereas the matching approach here does not depend on the support of the outcome, though the bias correction may perform better on continuous outcomes.

The second approach, inverse probability of treatment weighting, leverages correctly specified models for the propensity score of A_{i1} and A_{i2} to estimate the (direct) effect of A_{i1} (Murphy et al., 2001; Robins, 1998; Robins et al., 2000). Unfortunately, in practice, this approach can have poor performance due to unstable weights when the probability of $A_{i1} = 1$ is close to 0 or 1, which can be compounded by model misspecification (Goetgeluk et al., 2008). In these situations, matching will also lead to poor matches (in terms of covariate balance), but our bias correction approach may help mitigate this issue. Covariate-balancing propensity scores, which we include in our simulations below, were developed to improve the performance of inverse probability weighting under model misspecification and provide a good comparison to our approach

here (Imai & Ratkovic, 2015). In addition, a host of *doubly robust* methods have been developed that combine features of the structural nested and weighting approaches (Bang & Robins, 2005; van der Laan & Gruber, 2012). These methods require models for both (a) the outcome-covariate relationship and (b) the 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. Telescope matching possesses an approximate double robustness property in the sense that if either (a) the matching is exact or close to exact, or (b) the bias correction model is correctly specified, then the resulting telescope matching estimator will generally have small bias. This differs from the traditional notion of double robustness since that relies on the consistency of at least one of two models, whereas approximate double robustness here relies on finite-sample quality of the matching.

4 | SIMULATION STUDY

We evaluate the performance of telescope matching against existing direct effect methods using a simulation in which we artificially introduce model misspecification, following an approach similar to that of Kang and Schafer (2007). Our assumed data generating process reflects a common situation encountered by researchers where there are many observed covariates with varying magnitudes of confounding. We have twenty observed baseline confounders, X_{i1}, X_{i2}, …, X_{i20}, that are each $\mathcal{N}(0, 1)$, and fifteen intermediate confounders Z_{i1}, Z_{i2}, …, Z_{i15}. The early treatment assignment, A_{i1}, follows $\text{Bern}(\pi_i)$, where $\text{logit}(\pi_i) = \sum_{j=1}^{20} j^{-1}(-1)^j X_{ij}$. Each of the fifteen intermediate confounders, Z_{ik}, is a function of early treatment and of another, unobserved, confounding factor affecting both Z_{ik} and outcome Y_i. Therefore, while each Z_{ik} is causally affected by early treatment A_{i1}, 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_{ik}. Each intermediate confounder is generated as $Z_{ik} = \frac{1}{2 \times k} A_{i1} + \gamma_{ik} + \frac{\delta}{k} U_{ik}$, where $\gamma_{ik} \sim \mathcal{N}(0, 1), U_{ik} \sim \mathcal{N}(0, 0.2)$ with $\gamma_{ik} \perp\!\!\!\perp U_{ik}$. 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 τ when conditioning on Z_{ik} in a naive manner.

Late treatment A_{i2} follows $\text{Bern}(r_i)$, where

$$\text{logit}(r_i) = -1.5 + 0.5A_{i1} + \sum_{j=1}^{20} \left(\frac{(-1)^{j-1} 3}{2 + j} \right) X_{ij} + \sum_{k=1}^{15} \left(\frac{3}{2 + k} \right) Z_{ik}.$$

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 $A_{i2} = 1$ to be between about 0.37 and 0.39 in our simulations depending on the magnitude of confounding. Finally, the outcome is generated as

$$Y_i = 210 + 27.4A_{i2} + \sum_{j=1}^{20} \left(\frac{13.7}{\lceil j/2 \rceil} \right) X_{ij} + \sum_{k=1}^{15} \left(\frac{\delta}{k} \right) U_{ik} + \epsilon_i$$

where $\epsilon_i \sim \mathcal{N}(0, 1)$ and δ is the same parameter that appears in the functional form of Z_{ik}. In this case, the effect of A_{i1} flows entirely through future treatment, so the true (direct) effect of early treatment is 0.

As in Kang and Schafer (2007), we simulate model misspecification by considering a scenario where each of the confounders are not measured directly but rather as one of three non-linear transformations.

$$\left(X^*_{i,\text{odd}}, X^*_{i,\text{even}}, Z^*_{ik}\right) = \left\{\exp(X_{i,\text{odd}}/2), (1+\exp(X_{i,\text{even}}))^{-1} + 10, (Z_{ik}/25 + 0.6)^3\right\}$$

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 of early treatment using several approaches: (1) a naive overspecified linear regression that conditions on both treatments and all confounders, baseline and intermediate; (2) a structural nested mean model that assumes the outcome model is linear and additive in all variables; (3) the sequential propensity score matching approach of Lechner (2004) (without any bias correction); (4) the doubly robust approach of Bang and Robins (2005) using the same outcome model as structural nested mean model and the propensity scores estimated as with the Lechner method; and (5) our telescope matching approach with the Mahalanobis distance metric, bias correcting with the same regression model as in the structural nested mean model. In our simulations, we set the number of units matched to each treated unit to $L = 3$. We also considered a standard (stabilized) inverse probability weighting estimator, and while it performs reasonably well in large samples given correct model specifications for both late treatment and the outcome, we omit it from the graphs

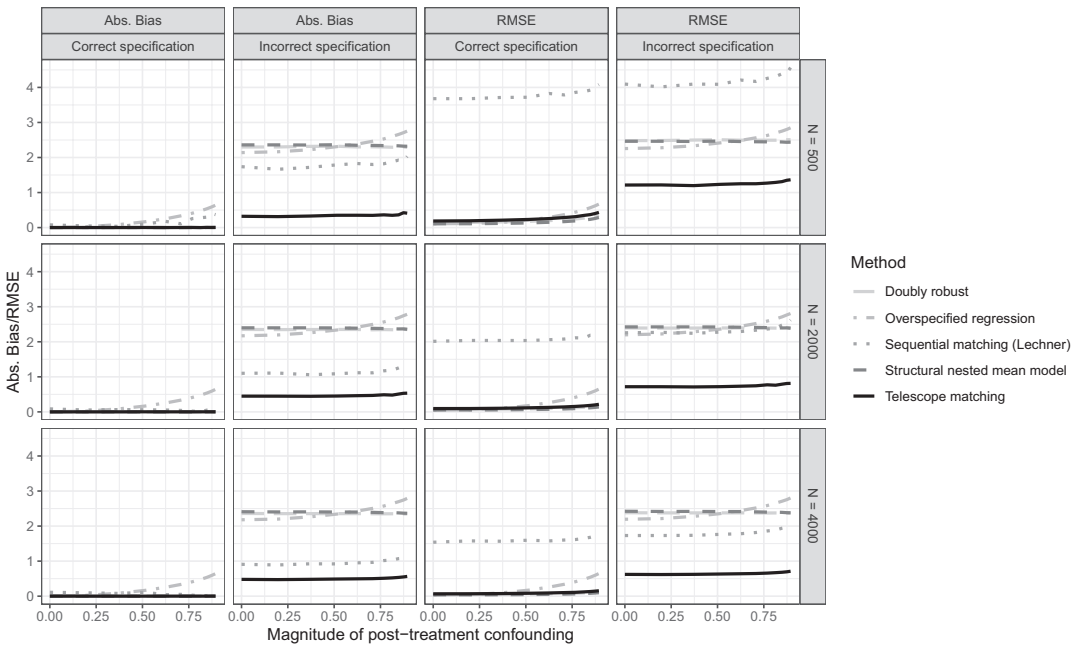

FIGURE 2 Performance of several estimators under simulated data with correct and misspecified models. The *y*-axis measures absolute bias (Abs. Bias) or root mean square error depending on the column

for expository reasons as the bias under misspecification is far larger than for any of the other methods, consistent with Kang and Schafer (2007). In addition, we evaluated the covariate-balancing propensity score method for marginal structural models which was designed to estimate weights in a way that is more robust to model misspecification (Imai & Ratkovic, 2015). Although the bias reduction under misspecification is considerable, the variance and RMSE far exceed the other methods and we also omit this from our graphs. In the Supplemental Materials, we show the figure with standard weighting results and covariate-balancing propensity scores (Figure SM.2) along with results on the confidence interval coverage of the various variance estimators described above (Figure SM.1).

Figure 2 plots the absolute value of the bias and the root mean squared error for all approaches under correct and incorrect model specifications. On the x-axis, we vary the amount of intermediate confounding (δ), which is translated into the partial correlation between U_{i1} and Y_i. This partial correlation is a deterministic function of δ: $0.2\delta(0.04\delta^2 + 1)^{-1/2}$. For each combination of parameter values, we carried out 10,000 iterations of our simulation. We find that both telescope matching and structural nested mean models are unbiased when the model is correctly specified, with the latter 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, the overspecified regression suffers from post-treatment bias, the magnitude of which grows as we increase the correlation between the intermediate confounder and Y_i. Lechner's sequential propensity score matching shows some bias that shrinks with sample size but grows slightly with the degree of intermediate confounding. The performance of the structural nested mean model and doubly robust estimator are nearly identical across all specifications.

When we introduce model misspecification, the performance of structural nested mean models is worse than that of telescope matching, with the gap growing as a function of the sample size. Telescope matching has considerably lower bias than all four of the other methods under misspecification and has the lowest root mean square error across all sample sizes and degree of post-treatment confounding. We also find that the root mean square error of telescope matching is decreasing in the sample size even with a relatively inflexible bias correction model. Finally, we find that the sequential propensity score matching has significantly higher bias under this misspecification even though it is a matching algorithm. This occurs because propensity score matching requires the correct specification of the propensity score model, whereas direct matching on covariates can be robust to those functional form assumptions.

Overall, the simulation results are promising for our proposed method. The findings 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 the other methods considered here, the bias of telescope matching can still be significant under the incorrect specification without a very large sample.

5 | EMPIRICAL ANALYSIS

Does negative advertising early in a campaign affect voter turnout or vote shares on election day? We apply the above methodology to a data set of Senate and gubernatorial elections in the United

States from 2000 until 2016. Our data expands on that used by Blackwell (2013) to estimate the effect of negativity by Democratic candidates on Democratic vote shares. In our analysis, we focus on the effect of incumbent negativity (regardless of party) on both voter turnout and vote shares. Voter turnout is defined as the percentage of citizen voting-age population (that is, those eligible to vote) who cast a ballot in the election, whereas vote share is the percent of the two-party vote for the incumbent. The data on advertising comes from the Wisconsin Advertising Project (Goldstein & Rivlin, 2007) and its successor the Wesleyan Media Project (Fowler et al., 2019), both of which code each political television advertisement as negative (mentioning the opponent) or positive (focusing entirely on the sponsor of the ad). Furthermore, these projects collect information on when these ads were shown, allowing us to create a measure of candidate negativity for the early part of the race versus the late part of the race. We focus on the effect of the incumbent candidate's decision to go negative and code $A_{i1} = 1$ if incumbent i's proportion of negative ads exceeded 50% from the end of the primary until the end of September. We code $A_{i2} = 1$ similarly for the months of October and the first week of November. After removing races that had no incumbent, no opponent, or had no television ads, we have a total of 144 races for the US Senate and 54 for state Governor. The median length of these campaigns—from the primary to the general election—is 21 weeks.

We estimate the direct effect of early negativity fixing late campaign tone to be negative—that is we seek to compare $Y_i(1, 1)$ to $Y_i(0, 1)$—since there are relatively few campaigns that switched from negative early to positive later, as seen in Table 1. To estimate these effects, we control for a host of potential covariates that might confound the relationship between the decision to go negative and the eventual outcomes. For baseline covariates, we include the length of the campaign in weeks; an indicator for whether the incumbent was a Democrat; the average support for the incumbent in baseline polling; the average percent undecided in baseline polling; the total number of ads shown by the major party candidates in the primary; an indicator for midterm versus presidential election year; a linear term for election year; baseline contributions to both candidates; the number of eligible voters in the state; and an indicator for office type. For intermediate covariates, we include several covariates measured at the beginning of the late-campaign period: the average support for the incumbent in polls; the average percent undecided; the log of the total number of ads shown through the early period; total number of contributions to either candidate through the end of the early period; and the average negativity of the challenger through the early part of the race. These intermediate confounders could at least plausibly be affected by early negativity by the incumbent.

We explore three different methods for estimating these direct effects. First, we simply estimate an overspecified linear regression with both treatments and both sets of covariates. Second, we estimate the effect using a linear structural nested mean model, where the covariates are modelled in the same way as the overspecified regression. Finally, we use telescope matching using the same linear regression models as the structural nested mean models for the bias correction. For telescope matching, we use $L = 3$ matches in both stages after checking that balance did not change dramatically with smaller matching ratios. In the Supplemental Materials, we present

TABLE 1 Count of campaign treatment histories

	Positive late ($A_{i2} = 0$)	**Negative late ($A_{i2} = 1$)**
Positive early ($A_{i1} = 0$)	93	44
Negative early ($A_{i1} = 1$)	11	50

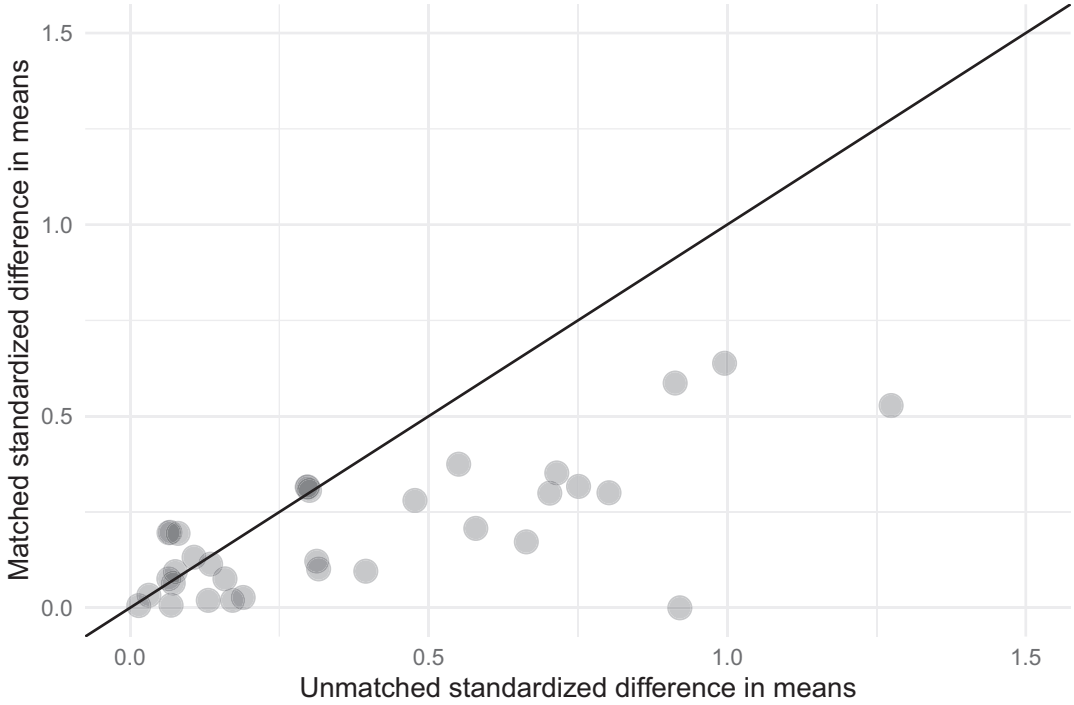

FIGURE 3 Comparison of balance before (*x*-axis) and after (*y*-axis) matching. Each dot represents a covariate its position represents the standardized difference in means across the treated and control groups in either period

a second empirical example from political science on the effect of membership in a labour union on racial attitudes of white Americans. That example has a larger sample size, but more extreme imbalance in the covariate distributions. In contrast to the campaign negativity example, we found that the estimated propensity scores in this example clustered very close to 0 and 1 as the treatment statuses were fairly well predicted by the covariates. While the larger sample size resulted in a greater reduction in post-matching imbalance due to the presence of more potential control matches, in both examples some residual imbalance remained due to insufficient control units in extreme parts of the covariate distribution.

5.1 | Findings

We begin by investigating how telescope matching addresses imbalances across both the early and late treatment covariate distributions. In Figure 3, we plot the standardized difference in means of each covariate in the matched and unmatched sample. For most of the covariates, there is a decrease in the imbalance across the treatment groups after matching (points below the 45-degree line), which we hope translates into robustness against model misspecification for those covariates. The few covariates that see increases in imbalance had fairly small imbalances in the unmatched sample and so the overall balance is much higher in the matched sample. In the Supplemental Materials, we present a full summary of the changes in balance for each variable along with an investigation of overlap in the propensity score distributions. Consistent with the nonparametric nature of the matching procedure, telescope matching reduces imbalance on

TABLE 2 Estimated effect of early incumbent negativity on voter turnout and incumbent vote percentage fixing late campaign tone to be negative ($N = 198$)

	Turnout			Vote share		
Method	**Est.**	**SE**	**95% Conf. Int.**	**Est.**	**SE**	**95% Conf. Int.**
Overspecified regression	0.827	1.283	(−1.687, 3.341)	−0.762	0.702	(−2.138, 0.614)
Structural nested mean model	2.284	1.072	(0.183, 4.386)	−1.926	0.758	(−3.411, −0.441)
Telescope matching	4.186	1.618	(1.015, 7.358)	−2.266	1.172	(−4.563, 0.031)

several squared terms of continuous covariates and the estimated propensity scores even though neither of these were used in the matching procedure. In spite of these improvements, there are still residual imbalances between the treated and control groups in both periods in terms of the propensity scores and other covariates. Our hope is that the bias correction step can adequately address these remaining imbalances, but there is the possibility that it drives some of the results we find below.

In Table 2, we present the estimated direct effects of early negativity on both voter turnout and the incumbent percentage of the vote. For voter turnout, each of the three methods produce a positive estimated effect of early negativity, though the estimate from telescope matching is much higher in magnitude than the other two estimates. This finding is interesting from a substantive perspective because campaign effects are usually thought to dissipate quickly due to recency bias (Gerber et al., 2011). The effects on incumbent vote shares, on the other hand, are negative and more consistent across telescope matching and structural nested mean models, though uncertainty is higher for the telescope matching estimator. Combined, these results are consistent with an account where early incumbent negativity actually mobilizes the challenger's supporters to turn out against the incumbent, leading to a backlash effect on vote shares.

Beyond the substantive results, this application demonstrates not only how telescope matching can be used to estimate the effects of time-varying treatments, but also how the choice of method can lead to dramatically different conclusions about these effects. In the Supplemental Materials, we discuss one likely source of this divergence due to model misspecification: when we add a handful of covariate interactions, the estimated effect of early negativity on turnout from the structural nested mean model becomes much closer to the estimate from telescope matching.

Finally, we note that both structural nested mean models and telescope matching rely on the sequential ignorability assumption, which is impossible to verify in observational settings like the present one. In our case, this assumption means that, the decision to go negative is independent of the potential outcomes of turnout and vote share conditional on the covariates listed above. This assumption will be most plausible when the covariates include all information that the campaigns used to make their decisions. And while our conditioning set does include many of those variables (campaign contribution, opponent's ad behaviour, polling), there may be private information that campaigns have to which we do not have access, making sequential ignorability less plausible. Thus, we believe an important avenue for future research is to extend the sensitivity analysis framework for matching (see, e.g., Rosenbaum, 1995, Chapter 4) to this sequential matching case.

6 | CONCLUSION

In this paper, we have introduced a novel method for estimating the effect of time-varying treatments. 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 structural nested mean models. This method will be useful to applied researchers who want to estimate the direct effect of early treatment 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. Third, it would be interesting to see how well this method generalizes to arbitrary time points and what steps that could be taken to mitigate potential power issues in that setting. In general, this paper illustrates how estimation of time-varying treatment 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.

ORCID

Matthew Blackwell http://orcid.org/0000-0002-3689-9527

REFERENCES

Abadie, A. & Imbens, G.W. (2006) Large sample properties of matching estimators for average treatment effects. *Econometrica*, 74(1), 235-267.

Abadie, A. & Imbens, G.W. (2008) On the failure of the bootstrap for matching estimators. *Econometrica*, 76(6), 1537–1557.

Abadie, A. & Imbens, G.W. (2011) Bias-corrected matching estimators for average treatment effects. *Journal of Business & Economic Statistics*, 29(1), 1–11.

Abadie, A. & Imbens, G.W. (2012) A martingale representation for matching estimators. *Journal of the American Statistical Association*, 107(498), 833–843.

Abadie, A. & Imbens, G.W. (2016) Matching on the estimated propensity score. *Econometrica*, 84(2), 781–807.

Bang, H. & Robins, J.M. (2005) Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61(4), 962–972.

Blackwell, M. (2013) A framework for dynamic causal inference in political science. *American Journal of Political Science*, 57(2), 504–520.

Dehejia, R.H. & Wahba, S. (1999) Causal effects in nonexperimental studies: reevaluating the evaluation of training programs. *Journal of the American Statistical Association*, 94(448), 1053–1062.

Fowler, E.F., Franz, M.M., Ridout, T.N. & Baum, L.M. (2019) Political advertising in 2016. The Wesleyan Media Project, Department of Government at Wesleyan University.

Gerber, A.S., Gimpel, J.G., Green, D.P. & Shaw, D.R. (2011) How large and long-lasting are the persuasive effects of televised campaign ads? Results from a randomized field experiment. *The American Political Science Review*, 105(1), 135–150.

Goetgeluk, S., Vansteelandt, S. & Goetghebeur, E. (2008) Estimation of controlled direct effects. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(5), 1049–1066.

Goldstein, K. & Rivlin, J. (2007) Congressional and gubernatorial advertising, 2003–2004. Combined File [dataset]. Final release.

Ho, D.E., Imai, K., King, G. & Stuart, E.A. (2006) Matching as Nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political Analysis*, 15(3), 199.

Huber, M., Lechner, M., & Strittmatter, A. (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, K. & Ratkovic, M. (2015) Robust estimation of inverse probability weights for marginal structural models. *Journal of the American Statistical Association*, 110(511), 1013–1023.

Imai, K., Keele, L. & Yamamoto, T. (2010) Identification, inference and sensitivity analysis for causal mediation effects. *Statistical Science*, 25(1), 51–71.

Imbens, G.W. (2004) Nonparametric estimation of average treatment effects under exogeneity: a review. *Review of Economics and Statistics*, 86(1), 4–29.

Jacobson, G.C. (2015) How do campaigns matter? *Annual Review of Political Science,* 18(1), 31–47.

Kang, J.D. & Schafer, J.L. (2007) Demystifying double robustness: a comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical Science*, *22*(4), 523–539.

King, G. & Nielsen, R. (2019) Why propensity scores should not be used for matching. *Political Analysis*, 27(4), 435–454.

van der Laan, M.J. & Gruber, S. (2012) Targeted minimum loss based estimation of causal effects of multiple time point interventions. *The International Journal of Biostatistics*, 8(1), Article 9.

Lau, R.R., Sigelman, L. & Rovner, I.B. (2007) The effects of negative political campaigns: a meta-analytic reassessment. *Journal of Politics*, 69(4), 1176–1209.

Lechner, M. (2004) Sequential matching estimation of dynamic causal models. IZA Discussion Papers 1042, Institute of Labor Economics (IZA).

Lechner, M. & Miquel, R. (2010) Identification of the effects of dynamic treatments by sequential conditional independence assumptions. *Empirical Economics*, 39(1), 111–137.

Murphy, S.A., van der Laan, M.J., Robins, J.M. & Conduct Problems Prevention Research Group (2001). Marginal mean models for dynamic regimes. *Journal of the American Statistical Association*, 96(456), 1410–1423.

Otsu, T. & Rai, Y. (2017) Bootstrap inference of matching estimators for average treatment effects. *Journal of the American Statistical Association*, 112(520), 1720–1732.

Richardson, T.S. & Rotnitzky, A. (2014) Causal etiology of the research of James M. Robins. *Statistical Science*, 29(4), 459–484.

Robins, J.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, J.M. (1997) Causal inference from complex longitudinal data. In: Berkane, M. (Ed.) *Latent variable modeling and applications to causality*, vol. 120 of Lecture notes in statistics. New York: Springer-Verlag, pp. 69–117.

Robins, J.M. (1998) Marginal structural models. In *1997 proceedings of the American statistical association section on Bayesian statistical science*, American Statistical Association, pp. 1–10.

Robins, J.M. (2000) Marginal structural models versus structural nested models as tools for causal inference. In: Halloran, M.E. & Berry, D. (Eds.) *Statistical models in epidemiology, the environment, and clinical trials*, vol. 116 of The IMA volumes in mathematics and its applications. New York: Springer-Verlag, pp. 95–134.

Robins, J.M. & Greenland, S. (1992) Identifiability and exchangeability for direct and indirect effects. *Epidemiology*, 3(2), 143–155.

Robins, J.M. & Rotnitzky, A. (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, J.M., Hernán, M.A. & Brumback, B.A. (2000) Marginal structural models and causal inference in epidemiology. *Epidemiology*, 11(5), 550–560.

Rosenbaum, P.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, P.R. (1995) *Observational studies*. New York: Springer-Verlag.

Rosenbaum, P.R. & Rubin, D.B. (1983) The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1), 41–55.

Rubin, D.B. (1974) Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5), 688.

Vansteelandt, S. (2010) Estimation of controlled direct effects on a dichotomous outcome using logistic structural direct effect models. *Biometrika*, 97(4), 921–934.

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

How to cite this article: Blackwell M, Strezhnev A. Telescope matching for reducing model dependence in the estimation of the effects of time-varying treatments: An application to negative advertising. *J R Stat Soc Series A*. 2021;00:1–23. https://doi.org/10.1111/rssa.12759