Robust inference for matching under rolling enrollment

Amanda K. Glazer and Samuel D. Pimentel*

June 22, 2023

Abstract

Matching in observational studies faces complications when units enroll in treatment on a rolling basis. While each treated unit has a specific time of entry into the study, control units each have many possible comparison, or "pseudo-treatment," times. Valid inference must account for correlations between repeated measures for a single unit, and researchers must decide how flexibly to match across time and units. We provide three important innovations. First, we introduce a new matched design, GroupMatch with instance replacement, allowing maximum flexibility in control selection. This new design searches over all possible comparison times for each treated-control pairing and is more amenable to analysis than past methods. Second, we propose a block bootstrap approach for inference in matched designs with rolling enrollment and demonstrate that it accounts properly for complex correlations across matched sets in our new design and several other contexts. Third, we develop a falsification test to detect violations of the timepoint agnosticism assumption, which is needed to permit flexible matching across time. We demonstrate the practical value of these tools via simulations and a case study of the impact of short-term injuries on batting performance in major league baseball.

Keywords: matching, block bootstrap, repeated measures, falsification test

1 Introduction

Matching methods attempt to estimate average causal effects by grouping each treated unit with one or more otherwise similar controls and using paired individuals to approximate the missing potential outcomes. Assuming that paired individuals are sufficiently similar on observed attributes and that no important unobserved attributes confound the comparison, the difference in outcomes approximates the impact of treatment for individuals in the pair

*Amanda K. Glazer is a doctoral candidate and Samuel D. Pimentel is an Assistant Professor in the Statistics Department at University of California, Berkeley, 367 Evans Hall, Berkeley, CA 94720. Correspondence should be addressed to amandaglazer@berkeley.edu.

[1]. Despite matching's transparency and intuitive appeal, it faces complications in datasets containing repeated measures for the same individuals over time. When only a single time of treatment is present, the primary challenge is deciding how to construct matching distances from pre-treatment repeated measures and assess outcomes using post-treatment repeated measures [2]. The situation is more complex under rolling enrollment, or staggered adoption, when individuals opt into treatment at different times [3]. Several authors [4; 5; 6; 7] proceed by matching each treated unit to the version of the control unit present in the data at the time of treatment. For example, in Imai et al. [7]'s reanalysis of data from Acemoglu et al. [8] on the impact of democratization on economic growth, countries undergoing democratizing political reforms are matched to similar control countries not undergoing such reforms in the same year.

Although this method is logical whenever strong time trends are present, in other cases it may overemphasize similarity on time at the expense of other variables. Bohl et al. [9] study the impact of serious falls on subsequent healthcare expenditures for elderly adults using patient data from a large healthcare system. While patients who fall could be matched to patients who appear similar based on recent health history on the calendar date of the fall, the degree of similarity in health histories is likely much more important than the similarity of the exact date at which each patient is measured. Following this idea, the GroupMatch algorithm [10] constructs matches optimally across time, prioritizing matching on important covariates over ensuring that units are compared at the same point in time.

Another example where rolling enrollment arises is in major league baseball (MLB). Quantifying the impact of injury on player performance in professional sports is important for both managers and players themselves. Increasingly, players are valued and compensated in a manner driven by quantitative metrics of past performance, but injuries have potential to disrupt the continuity between past and future performance [11; 12; 13; 14]. One way to quantify impact in this setting is as the difference between the value of a performance metric the player would have achieved in the absence of injury and the value of the same metric

achieved after a given injury. This quantity can be estimated using matching. However, players do not all get injured at the same time, so GroupMatch is a natural fit here. It allows us to match injured and non-injured players flexibly across time, because we likely do not care whether a player injured on June 1 is matched to a non-injured player on exactly the same day or a couple weeks earlier, e.g., May 15, so long as those players are sufficiently similar on other covariates such as recent performance.

Several challenges remain outstanding for matching methods under rolling enrollment. GroupMatch's flexible approach relies heavily on a strong assumption that time itself is not a confounder, and discussion of checking this assumption has been minimal so far. Even when flexible matching is warranted, the presence of multiple copies of the same control individual necessitates a constraint to ensure that a treated unit is not simply paired to multiple slightly different copies of the same control; several choices of this constraint exist permitting varying degrees of flexibility, and users must choose among them. Most importantly, for both GroupMatch and methods that match exactly in time there is substantial ambiguity about how to conduct valid inference. When multiple copies of a control individual are forbidden from appearing in the matched design, randomization inference may be used [5; 10] but no strong guarantees exist outside this special case.

In what follows, we present several innovations that greatly enhance the toolkit for matching and treatment effect evaluation under rolling enrollment. First, we introduce a new matched design called GroupMatch with instance replacement, which has computational, analytical, and statistical advantages over existing designs in many common settings. Second, we give a comprehensive characterization of a new block-bootstrap-based method for inference that applies broadly across existing methods for matching under rolling enrollment, including our new design. The block-bootstrap approach was originally suggested by Imai et al. [7] and is based on related work in the cross-sectional case by Otsu and Rai [15], but until now has not carried any formal guarantee. Finally, we introduce a falsification test to partially check the assumption of timepoint agnosticism underpinning GroupMatch's validity,

empowering investigators to extract evidence from the data about this key assumption prior to matching. We prove the validity of our bootstrap method under the most relevant set of constraints on reuse of controls, and we demonstrate the effectiveness of both the placebo test and the bootstrap inference approach through simulations and an analysis of injury data in major league baseball. In particular, the bootstrap method shows improved performance over linear-regression-based approaches to inference often applied in similar settings, while making much weaker assumptions.

The paper is organized as follows. Section 2 presents the basic statistical framework and reviews the GroupMatch framework, inference approaches for matching designs, and other related literature. In Section 3 we introduce a new constraint for use of controls in GroupMatch designs, leading to a new design called GroupMatch with instance replacement. Section 4 presents a block bootstrap inference approach for matching under rolling enrollment, and Section 5 evaluates it via simulation. In Section 6 we present a falsification test for the assumption that time is not a confounder. In Section 7 we apply our methods to evaluate whether minor injuries impact short-term MLB performance. Section 8 concludes.

2 Statistical framework

2.1 Setting and notation

We observe n subjects. For each subject i in the study, we observe repeated measures $(Y_{i,t}, \mathbf{X}_{i,t})$ for timepoints $t = 1, \ldots, T$, where $Y_{i,t}$ is an outcome of interestand $\mathbf{X}_{i,t}$ is a vector of covariates. We also observe a time of treatment initiation T_i for each subject, with $T_i \in \{1, \ldots, T\}$ for subjects who receive treatment at some point and $T_i = \infty$ for those who remain controls at all observed timepoints. We specify a burn-in period of length $L - 1$ during which no individuals are treated, i.e. $T_i \geq L$ for all i (or allow treatment at $t = 1$ by setting $L = 1$). We denote the collection of repeated measures for each subject i, along with T_i, as the trajectory O_i.

For clarity, we focus on "instantaneous" effects of treatment, with outcomes measured immediately following treatment at the same time when treatment is first initiated. Let $Y_{i,t}(0)$ be the potential outcome for unit i that would have been observed at time t if $T_i > t$, and let $Y_{i,t}(1)$ be the potential outcome that would have been observed if $T_i = t$. The finite sample average effect of treatment on the treated (ATT) is denoted by Δ:

$$\begin{aligned}\Delta &= \frac{1}{N_1}\sum_{i=1}^{N}\sum_{t=1}^{T} 1\{t = T_i\}[Y_{i,t}(1) - Y_{i,t}(0)] \\ &= \frac{1}{N_1}\sum_{i=1}^{N} D_i\left[Y_{i,t=T_i}(1) - Y_{i,t=T_i}(0)\right]\end{aligned}$$

Here the D_i variable is introduced as a convenient shorthand to indicate whether $T_i < \infty$. We assume that trajectories O_i are sampled independently from some infinite population, although we do not assume independence of observations within the same trajectory. Defining expectation $E(\cdot)$ with respect to sampling from this population, we define the population ATT as $\Delta_{pop} = E(\Delta)$. For future convenience, we also introduce a concise notation for conditional expectation (again, over the sampling distribution) of potential outcomes given no treatment through time t and the covariates observed in the previous L timepoints:

$$\mu_0^t(\mathbf{X}) = E[Y_{i,t}(0)|\{X_{i,t'}\}_{t'=t-L+1}^{t'=t} = \mathbf{X}, T_i > t]$$

Throughout, we abuse notation slightly by writing $\mu_0(\mathbf{X}_{i,t})$ to indicate conditional expectation given the L lagged values of $\mathbf{X}_i$ directly preceding time t, inclusive.

The potential outcomes framework adopted here represents one of many possible framings for studies with rolling enrollment. Pimentel et al. [10] define potential outcomes as functions of the length of time since treatment initiation, while both Ben-Michael et al. [3] and Athey & Imbens [16] define them as functions of the time of treatment initiation for the subject in question. In principle these alternate constructions are much richer than ours, allowing for much more general and complicated patterns of effects, but in practice all these authors use

simplifying assumptions or focus on estimands that reduce attention to at most two potential outcomes of interest for each individual at each timepoint. For example, both Pimentel et al. [10] and Ben-Michael et al. [3] allow for treatment effects to be measured at some follow-up time postdating the time of treatment rather than focusing on instantaneous effects, but since the length of follow-up is fixed in advance only two potential outcomes ever need to be considered for each unit. Similarly, the "no anticipation" assumption of both Ben-Michael et al. [3] and Athey & Imbens [16] ensures that there is exactly one potential outcome of interest associated with the control condition for each individual, and the "invariance to history" assumption of Athey & Imbens [16] collapses distinctions among potential outcomes under treatment. As such, the results we present below extend easily to all the potential outcomes frameworks just described, making the appropriate substitutions for our $Y_{i,t}(1)$ and $Y_{i,t}(0)$. While the potential outcomes framework of Imai et al. [7] is much more general than all the previously-mentioned works in allowing subjects to revert from treatment back to control, our framing also extends easily to it in the special case when no exit from treatment is allowed.

2.2 Identification assumptions

Pimentel et al. [10] studied the following difference-in-means estimator in designs where each treated unit is matched to C control observations. $M_{it,jt'}$ is an indicator for whether subject i at time t has been matched to subject j at time t':

$$\hat{\Delta} = \frac{1}{N_1} \sum_{i=1}^{n} D_i [Y_{i,t=T_i} - \frac{1}{C} \sum_{j=1}^{N} \sum_{t'=1}^{T} M_{iT_i,jt'} Y_{j,t'}]$$

Pimentel et al. [10] show that this estimator is unbiased for the population ATT under the following conditions:

1. Exact matching: matched units share identical values for covariates in the L timepoints preceding treatment.

2. L-ignorability: conditional on the covariate history over the previous L timepoints and the absence of treatment prior to baseline, an individual's potential outcome at a given time is independent of the individual's overall treatment status. Formally,

$$\{T_i < \infty\} \perp\!\!\!\perp Y_{i,t}(0) | T_i > t-1, \{X_{i,s}\}_{s=t-L+1}^{t}, \forall i.$$

Intuitively, this assumption prevents unobserved confounding that makes potential outcomes for treated subjects systematically different from those that remain controls even after accounting for information from a baseline period.

3. Timepoint agnosticism: mean potential outcomes under control do not differ for any instances with identical covariate histories at different timepoints. Formally, for any set of L covariate values $\mathbf{X}$,

$$\mu_0^t(\mathbf{X}) = \mu_0^{t'}(\mathbf{X}) = \mu_0(\mathbf{X}) \text{ for any } 1 \leq t, t' \leq T.$$

This assumption ensures that matching across time is reasonable by ruling out time trends other than those captured by time-varying covariates. For clarity we drop the t superscript when discussing the conditional expectation $\mu_0(\mathbf{X})$ in what follows, with the exception of Section 6 where we temporarily consider failures of this assumption.

4. Covariate L-exogeneity: future covariates do not encode information about the potential outcome at time t given covariates and absence of treatment over the previous L timepoints. Formally,

$$(X_{i,1}, ..., X_{i,T}) \perp\!\!\!\perp Y_{i,t}(0) | T_i > t-1, \{X_{i,s}\}_{s=t-L+1}^{t}), \forall i.$$

Like time agnosticism, covariate L-exogeneity is important to justify considering past and future instances from a control trajectory as part of the matching procedure. If

future instances' covariates include or are correlated with past instances' outcomes, then we may indirectly match on study outcomes introducing bias into our estimation step [17; 18]. Covariate L-exogeneity ensures that future covariates are safe to consider during the design stage.

5. Overlap: given that a unit is not yet treated at time $t-1 \geq L$, the probability of entering treatment at the next time point is neither 0 nor 1 for any choice of covariates over the L timepoints at and preceding t.

$$0 < P(T_i = t \mid T_i > t-1, X_i^t, \ldots, X_i^{t-L+1}) < 1 \qquad \forall t > L$$

While not stated explicitly in Pimentel et al. [10], we note that the authors rely on an overlap assumption of this type in the proof of their main result.

The exact matching assumption is no longer needed for asymptotic identification of the population ATT if we modify the estimator by adding in a bias correction term. As in Otsu and Rai [15] and Abadie and Imbens [19], we first estimate the conditional mean function $\mu_0(\mathbf{X})$ of the potential outcomes and use this outcome regression to adjust each matched pair for residual differences in covariates not addressed by matching. As outlined in Abadie and Imbens [19], bias correction leads to asymptotic consistency under regularity conditions on the potential outcome mean estimator $\widehat{\mu}(\cdot)$ (for further discussion of regularity assumptions on $\widehat{\mu}_0(\cdot)$ see the proof of Theorem 1 in Section A of the supplemental appendix). Many authors have also documented benefits from adjusting matched designs using outcome models [20; 21]. The specific form of our bias-corrected (i.e. model-adjusted) estimator is as follows:

$$\hat{\Delta}_{adj} = \frac{1}{N_1} \sum_{i=1}^{n} D_i[(Y_{i,t=T_i} - \hat{\mu}_0(\mathbf{X}_{i,T_i})) - \frac{1}{C} \sum_{j=1}^{N} \sum_{t'=1}^{T} M_{iT_i,jt'}(Y_{j,t'} - \hat{\mu}_0(\mathbf{X}_{i,t'}))]$$

Datasets with many variables, especially continuous variables or variables with many

categories, ensure that exact matching is rarely possible in practice, and in light of this we focus primarily on estimator $\widehat{\Delta}_{adj}$ in what follows.

As discussed by Imai et al. [7] for settings without rolling enrollment, identification is possible under weaker assumptions if a difference-in-differences estimator is used instead of the difference-in-means. While we focus primarily on the simpler bias-corrected difference-in-means estimator for clarity of exposition, the difference-in-differences approach also offers advantages for our setting, and the new matching and inference strategies we propose extend naturally to such estimators. We provide further discussion in Section 4.3.

3 GroupMatch with instance replacement

Before discussing our method for inference in general matched designs under rolling enrollment, we introduce a new type of GroupMatch design. Pimentel et al. [10] described two different designs produced by GroupMatch denoted Problems A and B, designs we refer to as GroupMatch without replacement and GroupMatch with trajectory replacement respectively.

1. **GroupMatch without replacement**: each control unit can be matched to at most one treated unit. If a treated unit is matched to an instance of a control unit, no other treated unit can match to (any instance) of that control unit.

2. **GroupMatch with trajectory replacement**: each control *instance* can be matched to at most one treated unit. Each treated unit can match to no more than one instance from the same control trajectory. However, different treated units can match to different instances of the same control trajectory, so a single control trajectory can contribute multiple distinct instances to the design.

As our chosen names for these designs suggest, their relative costs and benefits reflect the choice between matching without and with replacement in cross-sectional settings. As

discussed by Hansen [22], matching without replacement (in which each control may appear in at most one matched set), leads to less similar matches compared to matching with replacement (in which controls can reappear in many matched sets) since in cases where two treated units both share the same nearest control only one can use it. On the other hand, matching without replacement frequently leads to estimators with lower variance than those from matching with replacement, where an individual control unit may appear in many matched sets, making the estimator more sensitive to random fluctuations in its response. Thus, one aspect of choosing between these designs is a choice about how to strike a bias-variance tradeoff. The other important aspect distinguishing these designs is that randomization inference, which is based on permuting treatment assignments in each matched set independently of others, generally requires matching without replacement. Specifically, when multiple controls may be matched to each treated unit and replacement is allowed, the resulting configuration of treated and control units no longer resembles the design of a blocked or matched experiment.

These same dynamics play out in comparing GroupMatch without replacement and GroupMatch with trajectory replacement. GroupMatch without replacement ensures that responses in distinct matched sets are statistically independent (under a model in which trajectories are sampled independently), allowing for randomization inference, and ensures that the total weight on observations from any one control trajectory can sum only to $1/C$, ensuring that the estimator's variance cannot be too highly inflated by a single trajectory with large weight. The resulting data configuration also resembles what might be obtained in a sequential experimental design employing matching-on-the-fly as discussed by Kapelner & Krieger [23] and Pimentel et al. [10], and inference may be conducted using the associated randomization distribution. On the other hand, GroupMatch with trajectory replacement leads to higher-quality matches and reduced bias in matched pairs, although overlap among the matched sets formed makes it difficult to envision a corresponding "target trial."

We suggest a third GroupMatch design which leans even further towards expanding the

potential control pool and reducing bias.

3. **GroupMatch with instance replacement**: Each treated unit can match to no more than one instance from the same control unit, but control instances can be matched to more than one treated unit.

GroupMatch with instance replacement is identical to GroupMatch without trajectory replacement except that it allows repetition of individual instances within the matched design as well as non-identical instances from the same trajectory. As such, it is guaranteed to produce higher-quality matches than GroupMatch without trajectory replacement, but may lead to higher-variance estimators since individual instances may receive weights larger than $1/C$. Figure 1 illustrates these three GroupMatch methods with a toy example that matches injured baseball players to non-injured players based on on-base percentage (OBP).

In practice we view GroupMatch with instance replacement as a more attractive approach than GroupMatch with trajectory replacement almost without exception. One reason is that while the true variance of estimators from GroupMatch with instance replacement may often exceed that of estimators from GroupMatch with trajectory replacement by a small amount, our recommended approach for *estimating* the variance and conducting inference are not able to capture this difference. As we describe in Section 4, in the absence of a specific parametric model for correlations within a trajectory, inference proceeds in a conservative manner by assuming arbitrarily high correlations within a trajectory (much like the clustered standard error adjustment in linear regression). Since the variance advantage for GroupMatch with trajectory replacement arises only when correlations between instances within a trajectory are lower than one, the estimation strategy is not able to take advantage of them. This disconnect means that GroupMatch with trajectory replacement will not generally lead to narrower empirical confidence intervals, much as variance gains associated with paired randomized trials relative to less-finely-stratified randomized trials may not translate into reduced variance estimates [24].

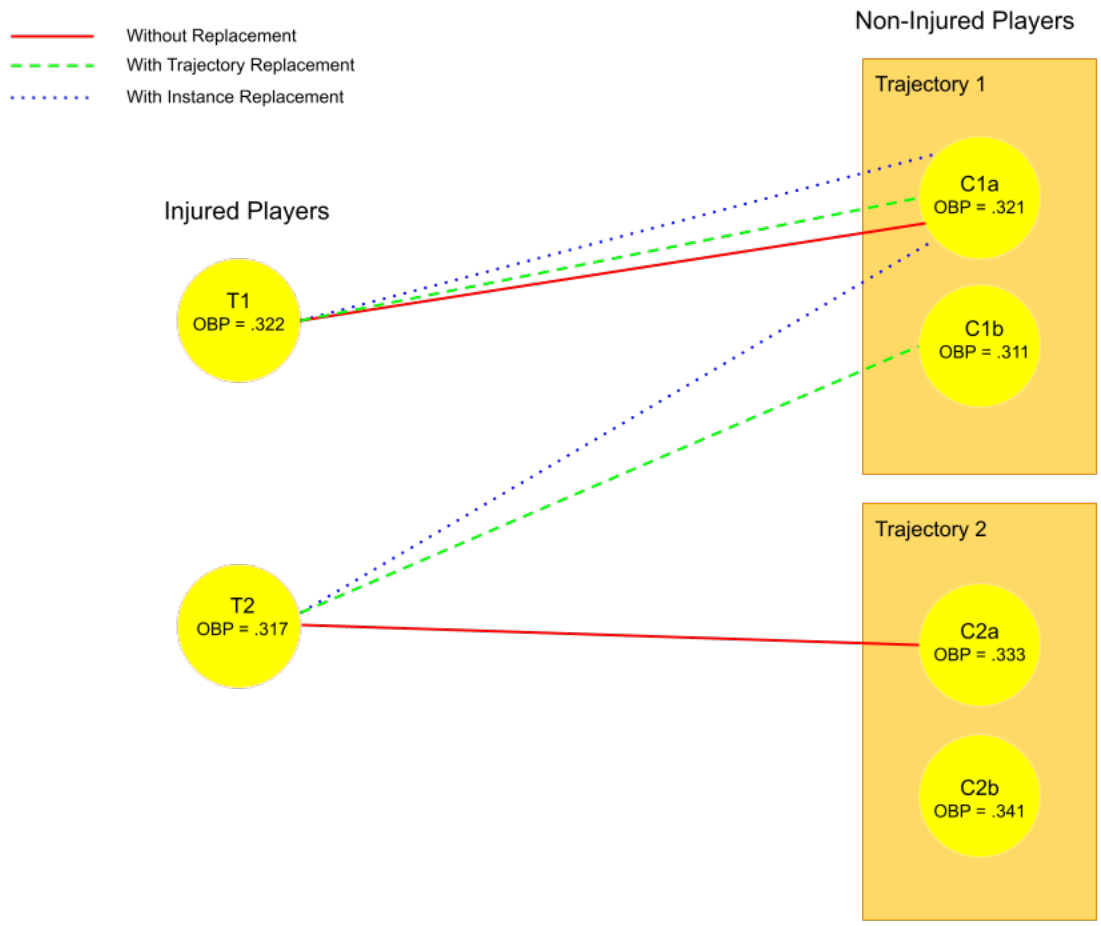

Figure 1: Toy example illustrating the three GroupMatch matching methods. Two injured baseball players (T1 and T2) are matched 1-1 to non-injured baseball players (C1a/b and C2a/b) based on player on-base percentage (OBP). Each non-injured player has two pseudo-injury times or instances. Under GroupMatch without replacement, T2 must match to an instance in Trajectory 2 because at most one instance from Trajectory 1 can participate in the match. Under GroupMatch with trajectory replacement, T2 can match to C1b but not to C1a, since multiple control instances can be chosen from the same trajectory as long as they are distinct. Under GroupMatch with instance replacement, both T1 and T2 are able to match to C1a. However, if each treated instance were matched to two control instances instead of one, GroupMatch with instance replacement would still forbid either T1 or T2 to match to a second instance in Trajectory 1.

A second important advantage of GroupMatch with instance replacement is its computational and analytical tractability relative to the other GroupMatch designs. One way to implement GroupMatch with instance replacement as a network flow optimization problem is to remove a set of constraints in Pimentel et al. [10]'s Network B (specifically the upper capacity on the directed edges connected to the sink node), and in Sections 5 and 7 we use this implementation for its convenient leveraging of the existing `groupmatch` package in R.

However, much more computationally efficient algorithms are also possible. Crucially, the removal of the constraint forbidding instance replacement means that matches can be calculated for each treated instance without reference to the choices made for other treated units; the C best matches for a given treated unit are simply the C nearest neighbor instances such that no two such control instances within the matched set come from the same trajectory. In principle, this allows for complete parallelization of the matching routine. On the analytical side, this aspect of the design makes it possible to characterize the matching algorithm as a generalized form of nearest neighbor matching, a strategy we adopt in the proof of Theorem 1 to leverage proof techniques used by Abadie and Imbens [25] for cross-sectional nearest neighbor matching. In light of these considerations, we focus primarily on GroupMatch with instance replacement in what follows, although the methods derived appear to perform well empirically for other GroupMatch designs too.

4 Block Bootstrap Inference

4.1 Inference methods for matched designs

Broadly speaking, there are two schools of thought in conducting inference for matched designs. One approach, spearheaded by Abadie and Imbens [25; 26; 19; 27], views the raw data as samples from an infinite population and demonstrates that estimators based on matched designs (which in this framework are considered to be random variables, as functions of random data) are asymptotically normal. Inferences are based on the asymptotic distributions of matched estimators. A second approach, described in detail in Rosenbaum [28; 29] and Fogarty [30], adopts the perspective of randomization inference in controlled experiments. Conditional on the structure of the match and the potential outcomes, the null distribution of a test statistic over all possible values of the treatment vector is obtained by permuting values of treatment within matched sets. When matches are exact and unobserved confounding is absent, strong finite sample guarantees hold for testing sharp null hypotheses

without further assumptions on outcome variables. Asymptotic guarantees for weak null hypotheses may be obtained too, assuming a sequence of successively larger finite populations [31]. Well-developed methods of sensitivity analysis are also available.

As described in Pimentel et al. [10], while standard methods of inference may be applied to GroupMatch without replacement, in which control individuals contribute at most one unit to any part of the match, none have been adequately developed for GroupMatch with trajectory replacement, in which distinct matched sets may contain different versions of the same control individual. For randomization inference, the barrier appears to be quite fundamental, because permutations of treatment within one matched set can no longer be considered independently for different matched sets. In GroupMatch with trajectory replacement, a treated unit receives treatment at one time and appears in a match only once; if treatment is permuted among members of a matched set so that a former control now attains treatment status, what is to be done about other versions of this control unit that are present in distinct matched sets? We note that similar issues arise when contemplating randomization inference for general cross-sectional matching designs with replacement, and we are aware of no solutions for randomization inference even in this simpler case.

In contrast, the primary issue in applying sampling-based inference to GroupMatch designs with trajectory replacement is the unknown correlation structure for repeated measures from a single control individual. The literature on matching with replacement provides estimators for pairs that are fully independent [27] and for cases in which a single observation appears identically in multiple pairs [25], but not for the intermediate case of GroupMatch with trajectory replacement where distinct but correlated observations appear in distinct matched sets. These issues extend beyond the GroupMatch family to any matched design under rolling enrollment in which control trajectories contribute to multiple matched sets, including those of Witman et al. [6] and Imai et al. [7].

In what follows we give formal guarantees for a sampling-based inference method appropriate for general matching designs under rolling enrollment suggested by Imai et al. [7],

which generalizes a recent proposal of Otsu and Rai [15] for valid sampling-based inference of cross-sectional matched studies using the bootstrap. Although the bootstrap often works well for matched designs without replacement [32], naïve applications of the bootstrap in matched designs with replacement have been shown to produce incorrect inferences as a consequence of the failure of certain regularity conditions [26]. Intuitively, if matching is performed after bootstrapping the original data, multiple copies of a treated unit will necessarily all match to the same control unit, creating a clumping effect not present in the original data. However, Otsu and Rai [15] arrived at an asymptotically valid bootstrap inference method for matching by bootstrapping weighted and bias-corrected functions of the original observations *after* matching rather than repeatedly matching from scratch in new bootstrap samples. We show that a similar bootstrap approach, applied to entire trajectories of repeated measures in a form of the block bootstrap, provides valid inference for matched designs under rolling enrollment. Note that in our formal results we focus on GroupMatch with instance replacement as the most difficult case, since the designs of Witman et al. [6] and Imai et al. [7] may be understood as restricted special cases in which matching on time is exact.

4.2 Block Bootstrap

In order to conduct inference under GroupMatch with trajectory or instance replacement we propose a weighted block bootstrap approach. We rearrange the GroupMatch ATT estimator from Section 2 as follows, letting $K_M(i,t)$ be the number of times the instance at trajectory i and time t is used as a match.

$$\begin{aligned}\hat{\Delta}_{adj} &= \frac{1}{N_1}\sum_{i=1}^{N} D_i[(Y_{i,T_i} - \hat{\mu}_0(\mathbf{X}_{i,T_i})) - \frac{1}{C}\sum_{j=1}^{N}\sum_{t'=1}^{T} M_{iT_i,jt'}(Y_{j,t'} - \hat{\mu}_0(\mathbf{X}_{i,t'}))] \\ &= \frac{1}{N_1}\sum_{i=1}^{N}\left[D_i\left(Y_{i,T_i} - \hat{\mu}_0(\mathbf{X}_{i,T_i})\right) - (1-D_i)\sum_{t=1}^{T}\frac{K_M(i,t)}{C}(Y_{i,t} - \hat{\mu}_0(\mathbf{X}_{i,t}))\right] = \frac{1}{N_1}\sum_{i=1}^{N}\hat{\Delta}_i.\end{aligned}$$

Because different instances of the same control unit are correlated, we resample the *trajectory*-level quantities $\widehat{\Delta}_i$ rather than the instance-level quantities. Since the $\widehat{\Delta}_i$ are functions of the $K_M(i,t)$ weights in the original match, we do not repeat the matching process within bootstrap samples. In particular, we proceed as follows:

1. Fit an outcome regression $\widehat{\mu}_0(\cdot)$ for outcomes based on covariates in the previous L timepoints using only control trajectories.

2. Match treated instances to control instances using GroupMatch with instance replacement. Calculate matching weights $K_M(i,t)$ equal to the number of times the instance at time t in trajectory i appears in the matched design.

3. Calculate the model-adjusted ATT estimator $\widehat{\Delta}_{adj}$.

4. Repeat B times:

   (a) Randomly sample N elements $\widehat{\Delta}_i$ with replacement from $\{\widehat{\Delta}_1, \ldots, \widehat{\Delta}_N\}$.

   (b) Calculate the bootstrap bias-corrected ATT estimator $\widehat{\Delta}^*_{adj}$ for this sample of trajectories as follows:

$$\widehat{\Delta}^*_{adj} = \frac{1}{N_1} \sum_{i=1}^{N} \widehat{\Delta}^*_i$$

5. Construct a (1 - α) confidence interval based on the $\alpha/2$ and $1-\alpha/2$ percentile of the $\widehat{\Delta}^*_{adj}$- values calculated from the bootstrap samples.

This method is essentially a block bootstrap, very similar to the method proposed in Imai et al. [7]. Note that while the recipe above uses the nonparametric bootstrap, it may easily be generalized to other approaches such as the wild bootstrap and the Bayesian bootstrap. In particular, consider rewriting $\widehat{\Delta}^*_{adj}$ in terms of a new set of random variables $W^*_1, \ldots, W^*_N$ that we denote the bootstrap weights:

$$\widehat{\Delta}^*_{adj} = \frac{1}{N_1} \sum_{i=1}^{N} \widehat{\Delta}^*_i = \sum_{i=1}^{N} \frac{W^*_i}{\sqrt{N_1}} \widehat{\Delta}_i \tag{1}$$

To recover the nonparametric bootstrap, the bootstrap weights W_i^* are chosen as $Q_i/\sqrt{N_1}$, where Q_i is the number of times subject i is selected when sampling with replacement; if the W_i^* are chosen instead by sampling from a scaled Dirichlet or scaled two-point distribution, we obtain the Bayesian bootstrap and the wild bootstrap respectively (see Otsu & Rai [15] for specifics). To adapt the step-by-step algorithm for these approaches, we draw W_i^*s rather than $\widehat{\Delta}^*$s in step 4(a) and use (1) to calculate $\widehat{\Delta}^*_{adj}$ in step 4b.

Our main result below shows the asymptotic validity of this approach. Several assumptions, in addition to Assumptions 2-5 in Section 2.2, are needed to prove this result. We summarize these assumptions verbally here, deferring formal mathematical statements to Section A.1 of the supplemental appendix. First, we require the covariates X_i to be continuous with compact and convex support and a density both bounded and bounded away from zero. Second, we require that the conditional mean functions are smooth in $\mathbf{X}$, with bounded fourth moments. In addition, we require that conditional variances of the treated potential outcomes and conditional variances of nontrivial linear combinations of control potential outcomes from the same trajectory are smooth and bounded away from zero. We also require that conditional fourth moments of potential outcomes under treatment and linear combinations of potential outcomes under control are uniformly bounded in the support of the covariates. Finally, we make additional assumptions related to the conditional outcome mean estimator $\widehat{\mu}_0(\cdot)$, specifically that the kLth derivative of the true conditional mean functions $\mu_1^t(\cdot)$ and $\mu_0(\cdot)$ exist and have finite suprema, and that the $\widehat{\mu}_(\cdot)$ converges to $\mu_0(\cdot)$ at a sufficiently fast rate. Finally, we require mild regularity conditions on the bootstrap weights W_i^*, easily satisfied by construction in the bootstrap approaches we have mentioned. To state the theorem, we also define

$$\sqrt{N_1}U^* = \frac{1}{\sqrt{N_1}}\sum_{i=1}^{N}\left(\widehat{\Delta}_i^* - D_i\widehat{\Delta}_{adj}\right) = \sum_{i=1}^{N} W_i^*(\hat{\Delta}_i - D_i\hat{\Delta}_{adj}).$$

Theorem 1. *Under assumptions M, W, and R presented in Section A.1 of the supplemental*

appendix,

$$sup_r|Pr\{\sqrt{N_1}U^* \leq r|(\mathbf{Y}, \mathbf{D}, \mathbf{X})\} - Pr\{\sqrt{N_1}(\hat{\Delta}_{adj} - \Delta) \leq r\}| \xrightarrow{p} 0$$

as $N \to \infty$ with fixed control:treated ratio C.

Our regularity assumptions on the data-generating process and the regression estimator $\widehat{\mu}_0(\cdot)$ are modeled closely on those of Abadie and Imbens [25] and later Otsu and Rai [15], and our proof technique is very similar to arguments in Otsu and Rai [15]. Briefly, U^* is decomposed into three terms which correspond to deviations of the potential outcome variables around their conditional means, approximation errors for $\widehat{\mu}_0(\mathbf{X})$ terms as estimates of $\mu_0(\mathbf{X})$ terms, and deviations of conditional average treatment effects $\mu_1^t(\mathbf{X}) - \mu_0(\mathbf{X})$ around the population ATT Δ. Regularity conditions on the data-generating process ensure that the conditional average treatment effects converge quickly to the population ATT. Regularity assumptions on the regression estimator, combined with bounds on the largest nearest-neighbor discrepancies in $\mathbf{X}$ vectors due originally to Abadie and Imbens [25] and adapted to the GroupMatch with instance replacement design, show that the deviation between $\widehat{\mu}_0(\cdot)$ and $\mu_0(\cdot)$ disappears at a fast rate. Finally, a central limit theorem applies to the deviations of the potential outcomes. For details, see Section A of the supplemental appendix.

4.3 Difference-in-Differences Estimator

While we have focused so far on the difference-in-means estimator, Imai et al. [7] recommend a difference-in-differences estimator for matched designs with rolling enrollment in the context of designs that match exactly on time. This estimator can be used under rolling

enrollment as well, taking the following form under bias correction:

$$\begin{aligned}\hat{\Delta}_{DiD} &= \frac{1}{N_1}\sum_{i=1}^{N} D_i[((Y_{i,T_i} - \hat{\mu}_0(\mathbf{X}_{i,T_i})) - (Y_{i,T_i-1} - \hat{\mu}_0(\mathbf{X}_{i,T_i-1})))- \\ &\quad \frac{1}{C}\sum_{j=1}^{N}\sum_{t'=1}^{T} M_{iT_i,jt'}((Y_{j,t'} - \hat{\mu}_0(\mathbf{X}_{i,t'})) - (Y_{j,t'-1} - \hat{\mu}_0(\mathbf{X}_{i,t'-1})))] \\ &= \frac{1}{N_1}\sum_{i=1}^{N} D_i[((Y_{i,t=T_i} - \hat{\mu}_0(\mathbf{X}_{i,T_i})) - (Y_{i,t=T_i-1} - \hat{\mu}_0(\mathbf{X}_{i,T_i-1})))- \\ &\quad (1-D_i)\sum_{t=1}^{T}\frac{K_M(i,t)}{C}((Y_{i,t} - \hat{\mu}_0(\mathbf{X}_{i,t})) - (Y_{i,t-1} - \hat{\mu}_0(\mathbf{X}_{i,t-1})))] = \frac{1}{N_1}\sum_{i=1}^{N}\hat{\Delta}_i^{DiD}\end{aligned}$$

This estimator requires L lags to be measured at time $T_i - 1$, so a burn-in period of length L rather than $L - 1$ is needed.

A key advantage of this bias-corrected differences-in-differences estimator is that it relies on different identification assumptions than the bias-corrected difference-in-means: essentially, any assumption previously made on the potential outcome $Y_{i,t}(0)$ must now hold instead only for the post-pre potential outcome difference $Y_{i,t}(0) - Y_{i,t-1}(0)$. The resulting assumptions tend to be substantively weaker. In particular, Imai et al. [7] highlight how the L-ignorability assumption can be replaced by a parallel trends assumption that requires only that post-pre differences in potential outcomes be conditionally independent of treatment, allowing for different unobserved outcome intercepts for different individuals. The time-agnosticism assumption also becomes weaker when formulated for outcome differences, allowing for a constant linear trend in potential outcome means rather than requiring them to be invariant to time conditional on covariates.

We can easily adapt the results of the previous section to show that the block bootstrap gives valid inference for the difference-in-difference estimator when these identification assumptions hold. The inference procedure simply requires bootstrapping the $\hat{\Delta}_i^{DiD}$ terms in place of the $\hat{\Delta}_i$s defined above. While the regularity conditions given for Theorem 1 suffice for the new estimator, the proof (as presented in Section A of the supplemental appendix)

requires mild modification to work for this difference-in-differences estimator. In particular, the variance estimators include additional covariance terms. For more details, see Section A.3 of the supplemental appendix.

5 Simulations

We now explore the performance of weighted block bootstrap inference via simulation. In particular, we investigate coverage and length of confidence intervals compared to those obtained by conducting parametric inference for weighted least squares estimators with and without cluster-robust error adjustment for controls from the same trajectory.

5.1 Data Generation

We generate eight covariates, four of them uniform across time for each individual i, (i.e., they take on the same value at every timepoint):

$$\begin{pmatrix} X_{3,i} \\ X_{4,i} \end{pmatrix} \sim N \left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0.7 \\ 0.7 & 01 \end{pmatrix} \right]$$

$X_{2,i} \sim N(0,1)$ for control units and $X_{2,i} \sim N(0.25,1)$ for treated units

X_1 is also uniform across time, but it is correlated with a time-varying covariate, X_5, so we will introduce it below. The correlations between covariates (X_3 and X_4, and X_1 and X_5) are calibrated to the correlations observed between covariates in the baseball example in Section 7 (i.e., height and weight have a correlation of approximately 0.7, and lag OBP and age have a correlation of approximately -0.4).

For treated units:

$$\begin{pmatrix} X_{1,i} \\ X_{5,i} \end{pmatrix} \sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & -0.4 \\ -0.4 & 01 \end{pmatrix} \right]$$

$$X_7, X_8 \sim N(0,1) \text{ and } X_6 \sim N(0.5,1)$$

Four of the covariates are time-varying for control units. For each control unit, three instances are generated from a random walk process to correlate their values across time. Formally, for instance t in trajectory i, covariate j is generated as follows:

$$\begin{pmatrix} X_{1,i} \\ X_{5,i,1} \end{pmatrix} \sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & -0.4 \\ -0.4 & 01 \end{pmatrix} \right]$$

$$X_{j,i,1} \sim N(0,1) \text{ for } j = 6,7,8$$

$$X_{j,i,t} = X_{j,i,(t-1)} + \epsilon_{j,i,(t-1)} \text{ for } t = 2,3$$

$$\epsilon_{j,i,1}, \epsilon_{j,i,2} \sim N(0, 0.5^2)$$

Fixing $a_L = log(1.25)$, $a_M = log(2)$, $a_H = log(4)$ and $a_{VH} = log(10)$, and drawing the $\epsilon_{i,t}$ terms independently from a standard normal distribution, we define out outcome as:

$$Y_{i,t} = a_L \sum_{j=1}^{4} X_{j,i,t} + a_{VH} X_{5,i,t} + a_M(X_{6,i,t} + X_{8,i,t}) + a_H(X_{7,i,t}) + \Delta D_i + \epsilon_{i,t} \qquad (2)$$

The outcome for a unit is correlated across time as it is generated from some time-varying covariates. Each simulation consists of 400 treated and 600 control individuals. We consider 1:2 matching. The true treatment effect, Δ, is 0.25.

We consider two alternative ways of generating the continuous outcome variable besides model (2). First, we add correlation to the error terms within trajectories. Specifically, the $\epsilon_{i,t}$s for a given trajectory i are generated from a normal distribution with mean 0 and covariance matrix with off diagonal values of 0.8. Second, in addition to the correlated

error terms, we square the $X_{2,i,t}$ term in the model, so it is no longer linear. We also run simulations with poor overlap. See Section C of the Supplemental Appendix for results and discussion of simulation performance under poor overlap.

We compare the bias-corrected block bootstrap approach outlined in Section 4, using a linear outcome model and a nonparametric bootstrap, to the confidence intervals obtained from weighted least squares (WLS) regression and WLS with clustered standard errors. We focus on the nonparametric bootstrap, as opposed to alternatives such as the wild bootstrap, because of its more common prevalence in practice; however, for comparisons between the wild bootstrap and the nonparametric showing almost equivalent performance in a generally similar setting see Otsu & Rai [15]. We choose to compare to WLS because this is commonly recommended in matching literature [33; 34]. However, Abadie and Spiess [35] pointed out that standard errors from regression may be incorrect due to dependencies among outcomes of matched units, and identified matching with replacement as a setting in which these dependencies are particularly difficult to correct for. Our simulation results suggest that these difficulties carry over into the case of repeated measures. It is worth noting that the standard functions in R used to compute WLS with matching weights such as lm and Zelig (which calls lm), compute biased standard error estimates in most settings. See Section B of the supplemental appendix for details.

5.2 Results

Tables 1 and 2 show the coverage and average 95% confidence interval (CI) length, respectively, of WLS regression, WLS regression with clustered standard errors, and bootstrap inference using our model-adjusted ATT estimator, for each of our three simulation settings under 10,000 simulations. As misspecification of the estimated linear outcome model increases the bootstrap method is substantially more robust (although under substantial misspecification the bias-corrected method also fails to achieve nominal coverage). While the bootstrap confidence intervals are generally slightly wider than the WLS and WLS cluster

confidence intervals, this is to be expected as the wider confidence intervals lead to improved coverage. In settings where strong scientific knowledge about the exact form of the outcome model is absent, the bootstrap approach appears more reliable than its chief competitors.

Coverage	WLS	WLS Cluster	Bootstrap Bias Corrected
Linear DGP	92.6%	94.4%	94.0%
Linear DGP, Correlated Errors	89.4%	91.7%	94.4%
Nonlinear DGP, Correlated Errors	83.3%	86.2%	89.8%

Table 1: Coverage of the WLS, WLS cluster and bootstrap bias corrected methods of inference for our three simulation set-ups.

Average CI Length	WLS	WLS Cluster	Bootstrap Bias Corrected
Linear DGP	0.25	0.27	0.27
Linear DGP, Correlated Errors	0.25	0.27	0.30
Nonlinear DGP, Correlated Errors	0.26	0.28	0.31

Table 2: Average 95% confidence interval length for the WLS, WLS cluster and bootstrap bias corrected methods of inference for our three simulation set-ups.

Results in tables 1 and 2 are for GroupMatch with instance replacement, however matching with trajectory replacement performed very similarly in our simulations. Computation time was similar for GroupMatch with instance replacement and with trajectory replacement. In principle, GroupMatch with instance replacement should be substantially faster, however in its current form GroupMatch does not implement the most computationally efficient algorithm for with instance replacement. Over 100 iterations, the average matching computation time was 4.63 seconds for matching with instance replacement and 4.71 seconds for matching with trajectory replacement. The average block bootstrap computation time was 2.51 seconds. Computation time was calculated on an Apple M1 Max 10-core CPU with 3.22 GHz processor and 64 GB RAM running on macOS Monterey.

6 Testing for Timepoint Agnosticism

The key advantage of GroupMatch relative to other matching techniques designed for rolling enrollment settings is its ability to consider and optimize over matches between units at

different timepoints, which leads to higher quality matches on lagged covariates. This advantage comes with a price in additional assumptions, notably the assumption of timepoint agnosticism. Timepoint agnosticism means that mean potential outcomes under control for any two individual timepoints in the data should be identical; in particular, this rules out time trends of any kind in the outcome model that cannot be explained by covariates in the prior L timepoints.

While in many applications scientific intuition about the data generating process suggests this assumption may be reasonable, it is essential that we consider any information contained in the observed data about whether it holds in a particular case. Accordingly, we present a falsification test for timepoint agnosticism. Falsification tests are tests "for treatment effects in places where the analyst knows they should not exist," [36] and are useful in a variety of settings in observational studies [37]. In particular, our test is designed to detect violations of timepoint agnosticism, or "treatment effects of time" when they should be absent; rejections indicate settings in which GroupMatch is not advisable and other rolling enrollment matching techniques that do not rely on timepoint agnosticism are likely more suitable. While failure to reject may not constitute proof positive of timepoint agnosticism's validity, it rules out gross violations, thereby limiting the potential for bias.

To test the timepoint agnosticism assumption we use *control-control time matching*: matching control units at different timepoints and testing if the average difference in outcomes between the two timepoint groups, conditional on relevant covariates, is significantly different from zero using a bootstrap test. Specifically, restricting attention to trajectories i from the control group, we select two timepoints t_0 and t_1 and match each instance at one timepoint to one at the other timepoint using the GroupMatch optimization routine, based on similarity of covariate histories over the previous L timepoints. Since this match compares instances at two fixed time points, any optimal method of matching without replacement may be used. One practical issue arises: GroupMatch and related matching routines expect one group to be designated "treated," all members of which are generally retained in the

match, and the other "control," some members of which will be included, but both matching groups are controls in this case. We label whichever of the two groups has fewer instances as treated; without loss of generality, we will assume there are fewer instances at time t_1 and use these instances as the reference group to be retained.

The test statistic for the falsification test is motivated by the ATT estimator in section 2.2. Let N_c be the total number of control units and let N_{t_1} be the number of control instances at time t_1. Let $\hat{\mu}_0^{t_0}$ be a bias correction model fit on our new control group (i.e., control instances at time t_0). In addition, let $D'_{it} = 1$ if unit i is present at time t. We define the test statistic as follows:

$$\begin{aligned}
\hat{\Delta}_{cc} &= \frac{1}{N_{t_1}} \sum_{i=1}^{N_c} D'_{it_1}((Y_{i,t=t_1} - \hat{\mu}_0^{t_0}(\mathbf{X}_{i,t=t_1})) - \sum_{j=1}^{N_c} M_{it_1,jt_0}(Y_{j,t=t_0} - \hat{\mu}_0^{t_0}(\mathbf{X}_{j,t=t_0}))) \\
&= \frac{1}{N_{t_1}} \sum_{i=1}^{N_c} [D'_{it_1}(Y_{i,t=t_1} - \hat{\mu}_0^{t_0}(\mathbf{X}_{i,t=t_1})) - D'_{it_0} K_M(i,t_0)(Y_{i,t=t_0} - \hat{\mu}_0^{t_0}(\mathbf{X}_{i,t=t_0}))] \\
&= \frac{1}{N_{t_1}} \sum_{i=1}^{N_c} \hat{\Delta}_{cc,i}
\end{aligned}$$

We use a bootstrap test to test the following null hypothesis, where $E_0^{t_1}\{\cdot\}$ indicates expectation over the distribution of the covariates in control instances at time t_1.

$$E_0^{t_1}\left\{\mu_0^{t_0}(\mathbf{X})\right\} = E_0^{t_1}\left\{\mu_0^{t_1}(\mathbf{X})\right\}$$

In words, this null hypothesis says that, accounting for differences in the covariate distribution at times 0 and 1, the difference in the average outcomes of control instances at the two timepoints is zero.

The test constructs a bootstrap confidence interval as in Section 3 and checks whether the interval covers 0. If the interval covers 0, the test fails to reject. In steps:

1. Label control instances from the first group of trajectories at timepoint t_1 the new

"treated" units, and control instances from the second group of trajectories at timepoint t_0 the new "control" units.

2. Fit a bias correction model on the new control units.

3. Match the new treated units to the new control units and calculate the test statistic.

4. Repeat B times:

   (a) Randomly resample N_c elements $\hat{\Delta}*_{cc,i}$ with replacement from $\{\hat{\Delta}_{cc,1}, ..., \hat{\Delta}_{cc,N_c}\}$

   (b) Calculate $\hat{\Delta}_{cc}$ on the resampled data.

5. Construct a $(1-\alpha)$ confidence interval based on the $\alpha/2$ and $1-\alpha/2$ percentile of the values calculated from the boostrap samples.

6. If this confidence interval covers 0, fail to reject the null hypothesis.

We choose to use a bootstrap test here in line with with our inference methods in previous sections. However, it is worth noting that a permutation test is also feasible here.

A key consideration for the falsification test is which timepoints to choose as t_0 and t_1. The choice of timepoint comparison depends largely on what a plausible time trend would be for the problem at hand. For example, if you suspect a linear time trend, it makes sense to look at the first and last timepoints. If the trend is linear, this test should have high power to detect a problem in moderate to large samples. If one is uncertain about the specific shape of the time trend that is most likely to occur and wants to test for all possible trends, we recommend testing each sequential pair of timepoints (i.e., timepoints 1 and 2, 2 and 3, 3 and 4, and so on) and using a multiplicity adjustment.

The falsification test is subject to several common criticisms levied at falsification tests, particularly their ineffectiveness in settings with low power. One possible approach is to reconfigure the test to assume violation of timepoint agnosticism as a null hypothesis and seek evidence in the data to reject it; Hartman and Hidalgo [38] recommend a similar change for falsification tests used to assess covariate balance, called *equivalence tests*.

The implementation of these modifications is fairly straightforward. First, we must define an equivalence range for our outcome variable: a set of values for which the difference is substantively inconsequential. Let ϵ_L and ϵ_U denote the lower and upper bounds under which the outcome variable is deemed equivalent. Hartman and Hidalgo [38] recommend using $\epsilon = \pm 0.36\sigma$ as a default when researchers are unsure of an appropriate equivalence region. Next, prior to step 4 in our falsification test one simply subtracts ϵ_L (and in a separate run ϵ_U) from all treated outcomes. If either one-sided test fails to reject the null, then the test fails.

See Section D of the supplemental appendix for simulations illustrating this method.

7 Application: Baseball Injuries

We study the impact of short-term injury on hitting performance in observational data from major league baseball (MLB) during 2013-2017. Quantitative studies of major league hitting performance [39] and of injury trends and impact in athletics [12] have been performed repeatedly, but only a few studies so far have evaluated the impact of injury on position players' hitting performance. These have focused on specific injury types, and have not found strong evidence that injury is associated with a decline in performance [11; 13; 14].

We use GroupMatch to match baseball players injured at certain times to similar players at other points in the season that were not injured. We evaluate whether players see a decline in offensive performance immediately after their return from injury. In contrast to other studies, we pool across injury types to see if there is a more general effect of short term injury on hitter performance.

7.1 Data and Methodology

We use publicly-available MLB player data from Retrosheet.org and injury data scraped from ProSportsTransactions.com for the years 2013-2017. Our dataset is composed of player

height, weight and age, quantities that remain constant over a single season of play, as well as on-base percentage (OBP), plate appearances (PAs) at different points in the season, and dates of short-term injuries, in which the player's team designated him for a 7-10 day stay on the team's official injured list, for each year. OBP is a common measure of hitter performance and is approximately equal to the number of times a player reaches base divided by their number of plate appearances.[1]

For each non-injured player, we generate three pseudo-injury dates evenly spaced over their PAs. In each season, we match injured players to four non-injured players. Matches were formed using GroupMatch with instance replacement, matching on age, weight, height, number of times previously injured, recent performance measured by OBP over the previous 100 PAs, and performance over the entire previous year as measured by end-of-year OBP after James-Stein shrinkage[2] We choose to shrink the OBP using James-Stein to limit the impact of sampling variability for players with a relatively small number of PAs the previous season [40].

Table 3 shows the balance for each of the covariates before and after matching. For each covariate, matching shrinks the standardized difference between the treated and control means. The balance achieved is not perfect, especially for the number of previous injuries. This underlines the importance of combining matching with bias-correction to clean up imbalances not removed by matching.

7.2 Results

We compare the results for bias-corrected block bootstrap inference, WLS, and WLS with clustered standard errors. The ATT estimates are positive (0.010), but the 95% confidence intervals cover zero for all methods, indicating that there is not strong evidence that short term injury impacts batter performance. We present the results for 2017 in Figure 2. Results

[1]OBP = (Hits + Walks + Hit By Pitch) / (At Bats + Walks + Hit by Pitch + Sacrifice Flies)

[2]See `https://chris-said.io/2017/05/03/empirical-bayes-for-multiple-sample-sizes/` for discussion of James-Stein shrinkage to estimators with variable sample sizes.

Variable	Treated Mean	Control Mean Before	Control Mean After	Standardized Difference Before	Standardized Difference After
Height	73.7	73.1	73.4	0.26	0.14
Weight	213	209	212	0.24	0.07
2016 OBP (JS Shrunk)	.324	.328	.323	-0.09	0.02
Lag OBP	.336	.341	.338	-0.07	-0.02
Birth Year	1988	1988	1988	-0.08	-0.06
Number Previous Injuries	2.73	1.91	2.16	0.30	0.21

Table 3: Balance table for MLB injury analysis before and after matching each injured player to four non-injured players.

from each of 2013 - 2016 were substantively the same, as were results obtained by pooling the matched data across years. The data pass the timepoint agnosticism test, comparing the first and last pseudo-injury dates. We chose to compare the first and last pseudo-injury dates, because we were most concerned about player performance degrading over the course of the entire season due to fatigue. We also perform equivalence tests using $\epsilon = \pm 0.02$, which our data also pass.

We could have also chosen to use a difference-in-differences estimator here, using the difference in performance right before and after the injury, or pseudo-injury, date as our outcome. Results are substantively the same for both estimators. This similarity is due to the close lag OBP matches GroupMatch produces for this example.

8 Discussion

The introduction of GroupMatch with instance replacement, a method for block bootstrap inference, and a test for timepoint agnosticism provide substantial new capabilities for matching in settings with rolling enrollment. We now discuss a number of limitations and opportunities for improvement.

Our proof of the block bootstrap approach assumes the use of GroupMatch with instance replacement. The large-sample properties of matched-pair discrepancies are substantially easier to analyze mathematically in this setting than GroupMatch with trajectory replace-

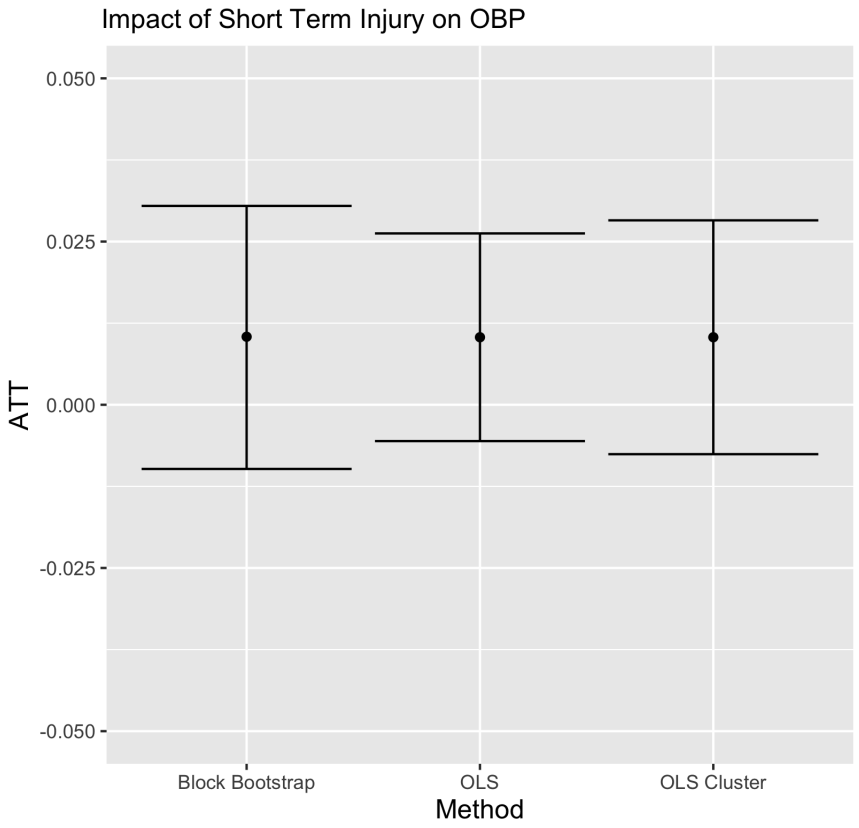

Figure 2: Estimates and 95% confidence intervals for block bootstrap, WLS and cluster WLS inference methods for the ATT in our 2017 baseball injury analysis.

ment or GroupMatch without replacement, designs in which different treated units may compete for the same control units, and the technical argument must be altered to account for this complexity. However, Abadie and Imbens [27] successfully characterized similar large-sample properties in cross-sectional settings for matching without replacement. While beyond the scope of our work here, we believe it is likely that this approach could provide an avenue for extending Theorem 1 to cover the other two GroupMatch designs. Empirically, we have found that the block bootstrap performs well when matches are calculated using any of the three GroupMatch designs.

Setting aside the technical barriers associated with extending the theory to GroupMatch without replacement, our new approach provides a competitor method to the existing randomization inference framework described by Pimentel et al. [10] available for GroupMatch without replacement. The randomization inference framework offers the advantage of closely-

related methods of sensitivity analysis and freedom from making assumptions about the sampling distribution of the response variables; on the other hand, the block bootstrap method avoids the need to assume a sharp null hypothesis. In general these same considerations arise in choosing between sampling-based inference and randomization-based inference for a cross-sectional matched study, although such choices have received surprisingly little direct and practical attention in the literature thus far.

As described in Section 6, the falsification test faces several criticisms, such as ineffectiveness in lower power settings. The modifications to construct equivalence tests as in Hartman and Hildago [38] address these concerns. However, even in the absence of such a change the falsification test may prove useful in concert with a sensitivity analysis. Sensitivity analysis, already widely studied in causal inference as a way to assess the role of ignorability assumptions, places a nonzero bound on the degree of violation of an assumption and reinterprets the study's results under this bound, often repeating the process for larger and larger values of the bound to gain insight. Such a procedure, which focuses primarily on assessing the impact of small or bounded violations of an assumption, naturally complements our falsification test, which can successfully rule out large violations but is more equivocal about minor violations.

Unfortunately, no sensitivity analysis appropriate for block bootstrap inference has been developed yet, either for timepoint agnosticism or other strong assumptions such as ignorability. The many existing methods for sensitivity analysis (developed primarily with ignorability assumptions in mind) are unsatisfying in our framework for a variety of reasons: some rely on randomization inference [29], others focus on weighting methods rather than matching [41; 42], and others are limited to specific outcome measures [43] or specific test statistics [44]. We view the development of compelling sensitivity analysis approaches to be an especially important methodological objective for matching under rolling enrollment.

Finally, we note that in cross-sectional settings moderate imbalances like those observed after matching in the baseball study in Section 7 can often be removed by refining the match

to include calipers [45; 46] or balance constraints [47; 48; 49] on important variables. For computational reasons these constraints are difficult to implement and use in full generality for GroupMatch designs. For example, some balance constraints rely on network flow representations of the matching problem that are not immediately compatible with the network flow representation underpinning GroupMatch. Further work to consider how calipers and balance constraints can be elegantly incorporated will enhance GroupMatch's effectiveness in practice.

Acknowledgments

The major league baseball performance data was obtained from and is copyrighted by Retrosheet (`www.retrosheet.org`). We thank the author and maintainer of the GitHub repository `https://github.com/robotallie/baseball-injuries` for making the injury data easily available. We also thank Eli Ben-Michael, Peng Ding, Avi Feller, Lauren Forrow, Shirshendu Ganguly, and Jiaqi Li for helpful conversations and feedback.

Funding information

Amanda Glazer (DMS RTG #1745640) and Samuel D. Pimentel (#2142146) acknowledge support from the National Science Foundation.

Conflict of interest

Authors state no conflict of interest.

References

[1] Stuart E. Matching methods for causal inference: A review and a look forward. Statistical science: a review journal of the Institute of Mathematical Statistics. 2010;25.

[2] Haviland A, Nagin DS, Rosenbaum PR, Tremblay RE. Combining group-based trajectory modeling and propensity score matching for causal inferences in nonexperimental longitudinal data. Developmental psychology. 2008;44(2):422-36.

[3] Ben-Michael E, Feller A, Rothstein J. Synthetic controls and weighted event studies with staggered adoption. arXiv preprint arXiv:191203290. 2021.

[4] Li YP, Propert KJ, Rosenbaum PR. Balanced risk set matching. Journal of the American Statistical Association. 2001;96(455):870-82.

[5] Lu B. Propensity score matching with time-dependent covariates. Biometrics. 2005;61:721-8.

[6] Witman A, Beadles C, Liu Y, Larsen A, Kafali N, Gandhi S, et al. Comparison group selection in the presence of rolling entry for health services research: Rolling entry matching. Health services research. 2019;54:492-501.

[7] Imai K, Kim IS, Wang E. Matching methods for causal inference with time-series cross-section data. Harvard University; 2020.

[8] Acemoglu D, Naidu S, Restrepo P, Robinson JA. Democracy does cause growth. Journal of Political Economy. 2019;127(1):47-100.

[9] Bohl AA, Fishman PA, Ciol MA, Williams B, LoGerfo J, Phelan EA. A longitudinal analysis of total 3-year healthcare costs for older adults who experience a fall requiring medical care. Journal of the American Geriatrics Society. 2010;58(5):853-60.

[10] Pimentel SD, Forrow LV, Gellar J, Li J. Optimal matching approaches in health policy evaluations under rolling enrollment. Journal of the Royal Statistical Society: Series A (Statistics in Society). 2020;183:1411-35.

[11] Begly JP, Guss MS, Wolfson TS, Mahure SA, Rokito AS, Jazrawi LM. Performance outcomes after medial ulnar collateral ligament reconstruction in Major League Baseball positional players. Journal of Shoulder and Elbow Surgery. 2018;27:282-90.

[12] Conte S, Camp CL, Dines JS. Injury trends in Major League Baseball over 18 seasons: 1998-2015. Am J Orthop. 2016;45:116-23.

[13] Frangiamore SJ, Mannava S, Briggs KK, McNamara S, Philippon MJ. Career Length and Performance Among Professional Baseball Players Returning to Play After Hip Arthroscopy. The American Journal of Sports Medicine. 2018;46:2588–2593.

[14] Wasserman EB, Abar B, Shah MN, Wasserman D, Bazarian JJ. Concussions are associated with decreased batting performance among Major League Baseball players. The American Journal of Sports Medicine. 2015;43:1127-33.

[15] Otsu T, Rai Y. Bootstrap inference of matching estimators for average treatment effects. Journal of the American Statistical Association. 2017;112:1720-32.

[16] Athey S, Imbens GW. Design-based analysis in difference-in-differences settings with staggered adoption. Journal of Econometrics. 2022;226(1):62-79.

[17] Rosenbaum PR. The consequences of adjustment for a concomitant variable that has been affected by the treatment. Journal of the Royal Statistical Society: Series A (General). 1984;147(5):656-66.

[18] Hansen BB. The prognostic analogue of the propensity score. Biometrika. 2008;95(2):481-8.

[19] Abadie A, Imbens GW. Bias-corrected matching estimators for average treatment effects. Journal of Business & Economic Statistics. 2011;29:1 11.

[20] Rubin DB. Using Multivariate Matched Sampling and Regression Adjustment to Control Bias in Observational Studies. Journal of the American Statistical Association. 1979:318-28.

[21] Antonelli J, Cefalu M, Palmer N, Agniel D. Doubly robust matching estimators for high dimensional confounding adjustment. Biometrics. 2018;74(4):1171-9.

[22] Hansen BB. Full matching in an observational study of coaching for the SAT. Journal of the American Statistical Association. 2004;99(467):609-18.

[23] Kapelner A, Krieger A. Matching on-the-fly: Sequential allocation with higher power and efficiency. Biometrics. 2014;70(2):378-88.

[24] Imbens GW. Experimental design for unit and cluster randomid trials. In: Conference International Initiative for Impact Evaluation, Cuernavaca; 2011. .

[25] Abadie A, Imbens GW. Large sample properties of matching estimators for average treatment effects. Econometrica. 2006;74:235-67.

[26] Abadie A, Imbens GW. On the failure of the bootstrap for matching estimators. Econometrica. 2008;76:1537-57.

[27] Abadie A, Imbens GW. A Martingale representation for matching estimators. Journal of the American Statistical Association. 2012;107:833 843.

[28] Rosenbaum PR. Covariance adjustment in randomized experiments and observational studies. Statistical Science. 2002;17(3):286-327.

[29] Rosenbaum PR. Observational Studies. New York, NY: Springer; 2002.

[30] Fogarty CB. Studentized sensitivity analysis for the sample average treatment effect in paired observational studies. Journal of the American Statistical Association. 2020;115(531):1518-30.

[31] Li X, Ding P. General forms of finite population central limit theorems with applications to causal inference. Journal of the American Statistical Association. 2017;112(520):1759-69.

[32] Austin PC, Small DS. The use of bootstrapping when using propensity-score matching without replacement: a simulation study. Statistics in medicine. 2014;33(24):4306-19.

[33] Ho DE, Imai K, King G, Stuart EA. Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. Political analysis. 2007;15:199-236.

[34] Stuart EA, King G, Imai K, Ho D. MatchIt: nonparametric preprocessing for parametric causal inference. Journal of statistical software. 2011.

[35] Abadie A, Spiess J. Robust post-matching inference. Journal of the American Statistical Association. 2021:1-13.

[36] Keele L. The statistics of causal inference: A view from political methodology. Political Analysis. 2015:313-35.

[37] Rosenbaum PR. Choice as an alternative to control in observational studies. Statistical science. 1999:259-78.

[38] Hartman E, Hidalgo FD. An Equivalence Approach to Balance and Placebo Tests. American Journal of Political Science. 2018;62(4):1000-13. Available from: `https://onlinelibrary.wiley.com/doi/abs/10.1111/ajps.12387`.

[39] Baumer BS. Why on-base percentage is a better indicator of future performance than batting average: An algebraic proof. Journal of Quantitative Analysis in Sports. 2008;4.

[40] Efron B, Morris C. Data analysis using Stein's estimator and its generalizations. Journal of the American Statistical Association. 1975;70:311-9.

[41] Zhao Q, Small DS, Bhattacharya BB. Sensitivity analysis for inverse probability weighting estimators via the percentile bootstrap. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2019.

[42] Soriano D, Ben-Michael E, Bickel PJ, Feller A, Pimentel SD. Interpretable sensitivity analysis for balancing weights. arXiv preprint arXiv:210213218. 2021.

[43] Ding P, VanderWeele TJ. Sensitivity analysis without assumptions. Epidemiology (Cambridge, Mass). 2016;27(3):368.

[44] Cinelli C, Hazlett C. Making sense of sensitivity: Extending omitted variable bias. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2020;82(1):39-67.

[45] Rosenbaum PR, Rubin DB. Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. The American Statistician. 1985;39(1):33-8.

[46] Yu R, Silber JH, Rosenbaum PR. Matching methods for observational studies derived from large administrative databases. Statistical Science. 2020;35(3):338-55.

[47] Rosenbaum PR, Ross RN, Silber JH. Minimum distance matched sampling with fine balance in an observational study of treatment for ovarian cancer. Journal of the American Statistical Association. 2007;102(477):75-83.

[48] Zubizarreta JR. Using mixed integer programming for matching in an observational study of kidney failure after surgery. Journal of the American Statistical Association. 2012;107(500):1360-71.

[49] Pimentel SD, Kelz RR, Silber JH, Rosenbaum PR. Large, sparse optimal matching with refined covariate balance in an observational study of the health outcomes produced by new surgeons. Journal of the American Statistical Association. 2015;110(510):515-27.