Title

Optimal matching approaches in health policy evaluations under rolling enrolment

Permalink
https://escholarship.org/uc/item/4930w26w

Journal
Journal of the Royal Statistical Society Series A (Statistics in Society), 183(4)

ISSN
0964-1998

Authors
Pimentel, Samuel D
Forrow, Lauren Vollmer
Gellar, Jonathan
et al.

Publication Date
2020-10-01

DOI
10.1111/rssa.12521

Peer reviewed

ORIGINAL ARTICLE

Journal of the Royal Statistical Society, Series A

Optimal matching approaches in health policy evaluations under rolling enrollment

Samuel D. Pimentel[1] | Lauren V. Forrow[2] | Jonathan Gellar[2] | Jiaqi Li[3]

[1] Department of Statistics, University of California, Berkeley, Berkeley, CA, 94720

[2] Mathematica Policy Research, Inc.

[3] Booz Allen Hamilton, Inc.

Correspondence
Samuel D. Pimentel, Department of Statistics, University of California, Berkeley, Berkeley, CA, 94720
Email: spi@berkeley.edu

Funding information
Not applicable

Comparison group selection is paramount for health policy evaluations, where randomization is seldom practicable. Rolling enrollment is common in these evaluations, introducing challenges for comparison group selection and inference. We propose a novel framework, GroupMatch, for comparison group selection under rolling enrollment, founded on the notion of time-agnosticism: two subjects with similar outcome trajectories but different enrollment periods may be more prognostically similar, and produce better inference if matched, than two subjects with the same enrollment period but different pre-enrollment trajectories. We articulate the conceptual advantages of this framework and demonstrate its efficacy in a simulation study and in an application to a study of the impact of falls in Medicare Advantage patients.

KEYWORDS

Matching, propensity score, rolling enrollment, network optimization, causal inference, health policy evaluation

1 | INTRODUCTION

1.1 | Rolling enrollment in health policy settings

Accurate estimation of causal effects is imperative to health policy evaluation and decision-making, but in large-scale evaluations of public health programs or policies, random assignment is seldom practicable. To evaluate the causal effect of these programs, quasi-experimental approaches such as propensity score matching (Rosenbaum and Rubin, 1983, 1985) are often used to construct a matched control group that appears similar to the treatment group to minimize nonrandom selection bias (Imbens and Wooldridge, 2009; Stuart, 2010). In recent Congressionally mandated evaluations of multi-billion dollar national health policies, including the Comprehensive Primary Care Initiative, the Bundled Payments for Care Improvement Initiative, and Quality Improvement Organizations, matching has been the canonical method of selecting comparable control groups (Chen et al., 2011; Peikes et al., 2018; Dummit et al., 2016).

Many health policy interventions do not have a fixed start date for all subjects; subjects enroll in the program on a rolling basis. For example, expecting mothers enrolled in Medicaid or Children's Health Insurance Program join CMS's Strong Start for Mothers and Newborns Initiative on or after they reach 12 weeks' gestation (Hill et al., 2018). Moreover, rolling enrollment is extremely common among people who receive care from home health agencies (HHA). These patients often experience a sudden decline in health status from acute events such as strokes or bone fractures from falling, or have disabilities and receive home health care services from HHAs as an alternative to nursing home care.

In longitudinal studies such as these, it is routine to match individuals using data from the pre-intervention period, also called the baseline period, to ensure that data used for matching do not reflect program effects. In interventions with rolling enrollment, the treatment group's baseline period is usually defined as some fixed time period immediately prior to the recorded date of entry into the intervention. For example, for a Chronic Disease Self-Management Program (CDSMP), the baseline period could be 6 months prior to the first day of attending the CDSMP course. Potential comparison subjects by definition lack a program enrollment date, so to obtain baseline period data for comparison group selection and later causal effect estimation we must assign pseudo-enrollment dates to potential comparison subjects. Assigning each potential comparison subject a series of pseudo-enrollment dates, rather than arbitrarily selecting a single date, allows for higher-quality matches by expanding the comparison pool to different temporal snapshots of each potential comparison subject. We call these different snapshots of the same potential comparison over time "versions" of the individual. However, this strategy introduces challenges for causal inference if more than one "version" of a given potential comparison subject is selected as part of the matched sample. Thus far, the matching literature has not fully resolved the tension between these important goals.

1.2 | Matching in settings with longitudinal measurements

Multivariate matching is a widely used technique for adjusting for confounding in observational studies. Matching is based on the intuitive idea of comparing a treated individual or individuals to a selected control individual or group of individuals who appear similar either in their probability of enrolling in the treatment, or in their potential outcomes under the control condition, or both (Hansen, 2008). Since the introduction of the propensity score (Rosenbaum and Rubin, 1983), the statistical literature has studied many aspects of matching techniques, as reviewed by Stuart (2010). However, only a small fraction of these methodological studies of matching focus on settings where the data include repeated measurements of covariates, outcomes, and (possibly) treatments for the same individuals over time.

Haviland et al. (2007) consider one such longitudinal setting in their study of the impact of joining a gang at age 14 on violent tendencies later in life, where each subject has repeated measurements for outcomes (violent tendencies), various covariates, and gang status at regular intervals. Trajectory models are used to identify important subgroupings of study subjects using the time-course data prior to treatment (in this case, from ages 11 to 13), and the subgroupings are then used to construct the match. But treatment itself is limited to a particular point in time (age 14), and its effect can be framed and discussed in the usual terms.

Li et al. (2001) consider a more complex setting where individuals wait differing lengths of time to receive treatment after enrolling in the study. Their recommended approach, called risk-set matching, involves matching individuals at the time they receive treatment to other individuals who have not yet received it, forming a group of matched sets distributed over time; this means that individuals who later receive treatment may enter the design as controls if they match to individuals receiving treatment earlier. As such, inference in this design measures the effect of a delay of treatment. Lu (2005) introduced a time-varying propensity score that computes the hazard of treatment at any timepoint as a function of time-dependent covariates at the same time point. Lu also discusses important differences between forming matched sets sequentially across time intervals and simultaneously matching sets for all time intervals at once; the latter design produces closer pairings but requires that time-dependent covariates obey an exogeneity condition (Hernán et al., 2001) lest access to post-treatment covariate information for the early-in-time matches induce bias. Risk-set matching can also be combined with carefully-chosen differential comparisons between treatments to construct artificial natural experiments, a design known as isolation (Zubizarreta et al., 2014, 2018).

In more recent work, Imai et al. (2019) consider matching in a setting with longitudinal data where treatment status may switch back and forth between treatment and control over time. Individuals with a particular contrast of patterns in past and present treatments are grouped and then further subdivided by covariate similarity. Inference is approached from a sampling perspective, rather than the randomization perspective used in Haviland et al. (2007), Li et al. (2001), and Lu (2005). As in risk-set matching, this framework allows for the same individuals to enter the match as treated units in one setting and controls in another; in fact, since this framework allows matching with replacement (in contrast to the other papers) the same individual may appear in both roles in the same design.

Finally, in recently published work Witman et al. (2019) present a method called rolling entry matching designed for the same health policy context we consider here. In this method treated units are compared to each control at the same point in calendar time. Similarity between units is evaluated using a single propensity score model fit to all data points based on their recent histories, and matching is done simultaneously for all treated units, either with or without replacement of control individuals.

These existing approaches differ from one another in many important respects. For example, Imai et al. (2019) assume observations are sampled from a larger population of interest while the other methods focus on finite-sample inference, and unlike the other methods it considers cases in which individuals cease treatment after beginning and uses controls multiple times across matched sets; Haviland et al. (2007) are unique in basing their matching design on estimates from a group-based trajectory model; and the risk-set matching papers, in contrast to the other methods, recommend conducting matching sequentially across time. However, they are all share one attribute of particular interest to us: some natural alignment of study subjects in time is assumed to be available. In Imai et al. (2019) and Witman et al. (2019), a common timescale is assumed among all individuals (e.g. measurements for one individual from December 2016 will be compared to measurements for another individual from December 2016). In Haviland et al. (2007) and Li et al. (2001), the alignment comes from measuring time past a certain individual-specific event; in the former, individuals are compared at the same age (i.e. time since birth), and in the latter, they are compared since their enrollment in the study and entry onto the waitlist. Our contribution to the literature is to relax this assumption of a fixed alignment in time and consider many possible alignments of control individuals for comparison with a given treated individual.

1.3 | GroupMatch

In this paper, we generalize the matching approaches for longitudinal data described above by contemplating settings in which concerns about exact alignment in time are secondary to concerns about balance on patterns of observed covariates. By exact alignment, we refer to the scenario where a treated individual who started treatment at time point t is matched to a comparison whose pseudo-start time is also t (Witman et al., 2019). We illustrate this issue in Figure 1, which depicts a treated subject T_1 who enrolls in treatment in November, having two emergency room (ER) visits in the prior two months (October and August). An otherwise-similar control subject in the same dataset, C_1, did not visit the emergency room at all in August through November; however, C_1 did have ER visits in July and May of the same year. We present four different "versions" of C_1 (C_{1a}, C_{1b}, C_{1c}, and C_{1d}), based on four different pseudo-enrollment dates. Under exact alignment, the only version of C_1 that would be available to match to T_1 is C_{1a}. However, our goal in forming a matched pair is to select individuals with very similar true propensities of enrollment in the treatment. If the pattern of ER visits over recent months is more strongly related to enrollment in the treatment group than calendar date, C_{1d} (which is defined by ER visits in the two months leading up to pseudo-enrollment) may be a more appropriate match to T_1 than C_{1a}. Existing matching methods from the literature, which require subjects to be arranged in some common alignment before individual pairwise matches are considered, cannot recognize or capitalize on this opportunity.

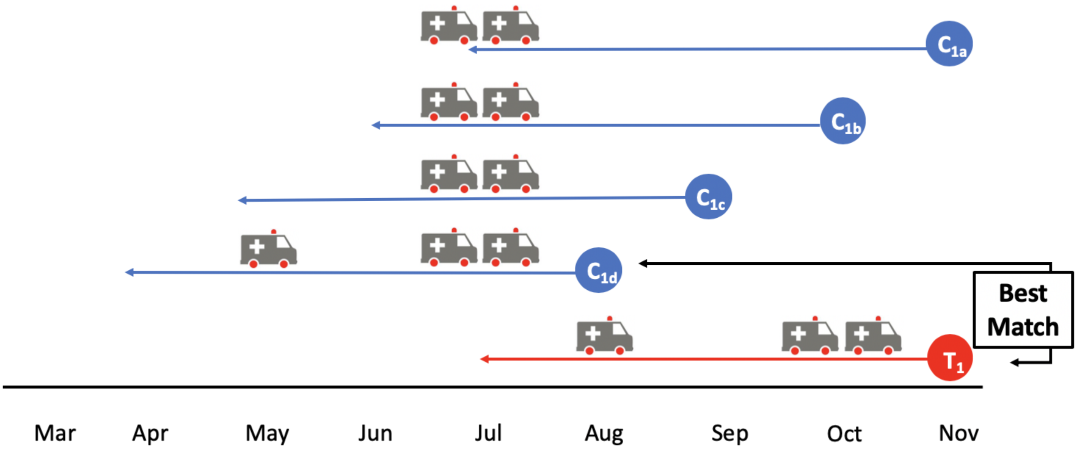

FIGURE 1 An illustration of matching of an treatment individual and different comparison "versions" with different pseudo-start dates. In this figure, each ambulance icon represents an ER visit. We consider a treatment intervention with a four-month baseline period, one treatment individual T_1 and four comparison versions $C_{1a}, C_{1b}, C_{1c}, C_{1d}$. If the pattern of ER visits in recent months is more important than the calendar date, then the best match for T_1 is C_{1d}.

In the example illustrated in Figure 1, we matched a treatment individual with a comparison individual whose start date is four months prior to that of the treatment individual. In practice, researchers can “go back" in time months or even years to find appropriate controls. The appropriate lag time depends on a few important factors, including availability of data, presence of secular trends, implementation of overlapping policy, and other exogenous factors that matching and subsequent identification methods cannot account for. For example, if we are interested in the effect of an opioid epidemic intervention on an uninsured population and a treated individual started to receive treatment in January 2015, we would only go back up to one year to look for appropriate controls as the Affordable Care Act Medicaid Expansion, which started in January 2014, has been shown to reduce opioid-related hospitalizations among uninsured individuals (Broaddus et al., 2018). For health policy studies, the dynamic nature of federal, state, and regional health policy interventions constrains the appropriate look-back period to no more than three years in most cases.

Relaxing the temporal alignment constraint is the key innovation in GroupMatch, our proposed solution to the rolling enrollment matching problem. With GroupMatch, we can consider matches across several different possible temporal alignments of a particular unique comparison subject and optimally select a single version to include in the final matched sample, improving covariate balance and treatment effect estimation relative to existing alternatives. We also consider a formulation of the problem in which different versions of the same unique comparison can be selected as matches for different treatment subjects. Both of these formulations are implemented in our R package, GroupMatch, which will soon be available on CRAN.

The rest of the paper is organized at follows. Section 2 discusses the statistical formulation of our problem, introduces important assumptions we call time-agnostic L-ignorability and covariate L-exogeneity, and demonstrates identification of an effect of treatment on the treated. In Section 3, we describe two GroupMatch algorithms for computing optimal matched designs under this statistical framework, each imposing different constraints on reuse of control subject information. We present a detailed simulation study in Section 4 which benchmarks GroupMatch against existing matching techniques. In Section 5 we apply GroupMatch to create a design for estimating the impact of falls on Medicare Advantage population enrolled in the Group Health Western Washington Integrated Group Practice. Section 6 discusses inference for effect estimates from GroupMatch. Section 7 concludes the paper with suggestions for future research directions.

2 | STATISTICAL FORMULATION

2.1 | Sampling framework and causal estimand

For each individual i in our study we observe a series of triples (Y_i^t, X_i^t, Z_i^t) for times $t = 1, \ldots, T$, where the Y_i^t is a scalar outcome, the X_i^t is a potentially vector-valued set of covariates, and Z_i^t is a nonnegative integer indicating time elapsed since enrollment in a treatment program of interest. We assume that individuals only enroll at most once in the treatment and remain enrolled, so for fixed i, Z_i^t is monotonically increasing in t, and strictly increasing in increments of one after enrollment. We call the collection $(\mathbf{Y}_i, \mathbf{X}_i, \mathbf{Z}_i)$ of observations for all t-values given some i the trajectory for individual i.

We assume that N_1 trajectories i are sampled i.i.d from a treated population of interest in which for all trajectories i we have $\{\exists t : Z_i^t > 0\}$, and that N_0 trajectories are sampled i.i.d. from a control population in which for all trajectories i we have $\{\forall t : Z_i^t = 0\}$. However, we allow for correlation between observations at different time points t, t' within the same individual i. Let $n = N_1 + N_0$ be the total number of distinct trajectories sampled.

Define the potential outcome $Y_i^t(z)$ for $z \in \{0, 1, \ldots, T\}$ as the outcome that would have been observed at time t for individual i if individual i had been enrolled for z timepoints at time t. Let $1 \leq P \leq T$ be a fixed length of follow-up. Define the finite sample average effect of the treatment on the treated (ATT) after follow-up as follows (where $1\{A\}$ is an indicator for event A):

$$\Delta = \frac{1}{N_1} \sum_{i=1}^{n} \sum_{t=1}^{T} 1\{Z_i^t = P\} \left[Y_i^t(P) - Y_i^t(0) \right]$$

We are primarily interested in estimating the population version of this effect, defined as follows. Note that expectation is with respect to the infinite treatment population from which treated trajectories are sampled:

$$\Delta_{pop} = E\left[\Delta\right].$$

From an empirical perspective, the population effect is of interest because in health policy settings the ultimate goal is generally to understand the potential impact of a policy on a broader population of eligible patients, not just the realized impact in the sample at hand. The assumption here that treated patients in the sample are drawn at random from this larger population of eligible patients is important to achieving population inference. For discussion of an alternate approach to inference in this setting focused only on the finite sample, see Section 6.

2.2 | Timepoint agnostic L-ignorability and covariate L-exogeneity

To identify the population ATT in observational data, we need some conditions to hold. Let L be some fixed number of time lags, and consider the following assumption:

$$Z_i^T \perp\!\!\!\perp Y_i^t(0) \,\Big|\, \left(Z_i^{t-P},\ \{X_i^s\}_{s=t-P-L}^{t-P}\right) \qquad \forall i.$$

This assumption, which we will call L-ignorability, says that an individual's potential outcome in the absence of treatment is independent of his/her present or future time in treatment, conditional on the covariate history over the baseline period (i.e. the L timepoints preceding the follow-up period of length P) and on any treatment enrollment in or prior to baseline. This assumption can be viewed as a weaker version of the causal exogeneity assumption of Hernán et al. (2001) since it enforces exogeneity only conditional on covariate history over L lags, or a stronger version of sequential ignorability (Robins, 2000) since it does not permit conditioning on past outcome values. L-ignorability also differs slightly from these other conditions in imposing assumptions only on $Y_i^t(0)$ rather than on a potential outcomes under both treatment and control conditions, which is possible here since we are interested in effects of treatment on the treated.

Notice that L-ignorability does not impose any restrictions on the longitudinal variation of treatment probability or potential outcomes; without additional assumptions it is not possible to pool information about either across time. While in certain settings such a strict approach may be warranted, in other situations such as those described in section 1.1 the exact timing of the enrollment is less important than the pattern of covariates that precedes it. In view of these considerations we propose an additional assumption. Define the mean of potential outcomes under control conditional on baseline covariates:

$$\mu_0^t(\mathbf{X}) = E\left[Y^t(0) \mid (X^{t-P}, \ldots, X^{t-P-L}) = \mathbf{X}, Z^{t-P} = 0\right]$$

We make the following assumption on the $\mu_0^t(\cdot)$ functions:

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

This condition, which we will call timepoint agnosticism, requires identical mean potential outcomes under control for any two individual-times in the data with treatment enrollment time less than P and identical lagged covariate histories, even if the timepoints differ. When the two assumptions hold simultaneously, a situation we denote by TALI, a much wider range of design choices may be available than L-ignorability alone would provide. Consider an exact matching approach where we seek to match each newly treated individual in our data to a control individual with an identical probability of treatment initiation. Under L-ignorability alone, we could only compare the lagged history of an individual who initiates treatment at $t - P$ to lagged covariate histories for controls over the period $t - P, \ldots, t - P - L$, and if no control with an identical lagged history in this period could be found then the treated unit could not contribute to the design. However, under TALI we would be free to consider any period over L timepoints of a control's history as a potential comparison for the treated unit, and an exact match on covariates might well exist in this much larger set of possible comparisons.

In addition to TALI, we will require that the covariate process be exogenous, in the sense of the causal exogeneity of Hernán et al. (2001). This assumption ensures both that future covariates do not encode information about earlier potential outcomes — lest we match on concomitant variables (Rosenbaum, 1984) — and that prior covariates more than L lags removed do not act as confounders. Specifically, assume the following, which we will call covariate L-exogeneity:

$$(X_i^1, \dots, X_i^T) \perp\!\!\!\perp Y_i^t(0) \,\Big|\, \left(Z_i^t,\ \{X_i^s\}_{s=t-P-L}^{t-P}\right) \qquad \forall i.$$

2.3 | Identification under exact matching

Given a dataset for which we believe the TALI and covariate L-exogeneity assumptions and in which controls are vastly more numerous than treated subjects, a natural design consists of matched sets, each containing one treated subject and several controls, all subjects in a set having similar covariate histories over a fixed-length “pre-enrollment" period without necessarily sharing the same enrollment or follow-up times. The motivation for creating matched sets rather than matched pairs is the opportunity to make more efficient use of the large control pool and reduce the standard error of the estimator.

In the sections that follow we describe strategies for forming matching designs in which the treated subject and matched control subjects do not necessarily share the same alignment in time, but are nonetheless as similar as possible (in a global sense across all matched sets) in their lagged histories. We now show that when TALI and covariate L-exogeneity hold and matching is exact on lagged histories, matched designs allow identification of the ATT without need for alignment in time within matched sets.

Assume we match each treated subject to w control observations without replacement, and let $M_{it,jt'}$ be an indicator for whether subject i at time t has been matched to subject j at time t' in our matched design. We will estimate the ATT by the difference in outcome means between the matched treated subjects and matched controls, which can be represented as follows:

$$\widehat{\Delta} = \frac{1}{N_1} \sum_{i=1}^{n} \sum_{it=1}^{T} 1\left\{Z_i^t = P\right\} \left[Y_i^t - \frac{1}{w} \sum_{j=1}^{N} \sum_{t'=1}^{T} M_{it,jt'} Y_j^{t'} \right].$$

Finally, we will call our matching design exact if $(X_i^{t-P}, \dots, X_i^{t-P-L}) = (X_j^{t'-P}, \dots, X_j^{t'-P-L})$ whenever $M_{it,jt'} = 1$, and uniformly exact if the ordered vector of all indicators $M_{it,jt'}$ is chosen uniformly at random from among all vectors associated with exact designs. The following result demonstrates that under TALI, covariate L-exogeneity, and uniformly exact matching, $\widehat{\Delta}$ is equal to the finite sample effect Δ plus a disturbance term D_n which has expectation zero in the population, so that $\widehat{\Delta}$ is unbiased for Δ_{pop}.

Theorem 1 *Suppose that TALI and covariate L-exogeneity hold, and that our matched design is uniformly exact. Then the* *following holds:*

$$\widehat{\Delta} = \Delta + D_n$$

where $E(D_n) = 0$ *when expectation is taken with respect to the infinite population, so that*

$$E(\widehat{\Delta}) = \Delta_{pop}.$$

We leave the proof of this proposition to the Appendix. Note that if matching is not exact, the decomposition of $\widehat{\Delta}$ contains an additional additive term. Abadie and Imbens (2006) refer to this term as conditional bias and provide conditions under which it converges asymptotically to zero for a variety of matching methods.

3 | COMPUTING MATCHES WITH GROUPMATCH

3.1 | Desiderata for rolling enrollment matching

As mentioned in Section 2, the freedom given by TALI to match individuals without exact agreement on follow-up times is an advantage that should permit higher quality matches by allowing treated subjects not just to choose among controls but among many different versions of controls over time. Such a design could be created simply by treating each unique combination of control subject and follow-up time as a separate individual and using standard fixed or variable ratio matching methods (Rosenbaum, 1989; Ming and Rosenbaum, 2001; Pimentel et al., 2015b) to match members of this newly-expanded control population to the treated subjects. However, this simple approach creates practical problems. Since copies of the same control subject with slightly different follow-up times will have highly correlated covariate histories, it is likely that the second- and third-best matched comparisons for a given treated subject will be slightly shifted-in-time versions of the best matched comparison. Thus a matching algorithm that intends to match multiple controls to a treated subject will instead effectively match only a single control.

In practice, to ensure that matched sets both take advantage of the flexibility to avoid exact time constraints and include information from a variety of different control subjects, we wish to constrain each matched set to include no more than one version of any control subject. While this constraint is simple to describe in words or in mathematical notation, it is more challenging to implement without incurring undue computational cost. Since computational efficiency relies on minimum-cost network flow optimization as described in Section 3.2, all constraints must be represented as restrictions of flow across a graph. Therefore, to guarantee that our one-version-per-control-per-matched-set constraint permits computation of matches at scale, we must reframe it as a network flow problem.

3.2 | Review: matching via network flow optimization

In the traditional minimum-distance pair matching problem, each treated subject among a sample of size N_1 and indexed by i must be paired to exactly one control subject from a sample of size N_0 (assumed to exceed N_1) and indexed by j in such a way that the sum of predefined pairwise covariate distances δ_{ij} for the selected pairs is as small as possible. Formally, we may write it as follows.

$$\begin{aligned}
&\min \sum_{i=1}^{N_1} \sum_{j=1}^{N_0} x_{ij}\delta_{ij} \\
&\text{s.t.} \\
&\sum_{j=1}^{N_0} x_{ij} = 1 \qquad \text{for all } i \in \{1, \dots, N_1\} \\
&\sum_{i=1}^{N_1} x_{ij} = 1 \qquad \text{for all } j \in \{1, \dots, N_0\} \\
&x_{ij} \in \{0, 1\} \qquad \text{for all } i \in \{1, \dots, N_1\}, j \in \{1, \dots, N_0\}
\end{aligned}$$

This specification of the problem requires pairs to be formed and does not permit any control to be used in multiple pairs (matching “without replacement"). These constraints can easily be modified to allow formation of sets with 1 treated unit and w controls (1:w matching) by changing the right-hand side of the first constraint to w, or to allow reuse of controls (matching “with replacement") by removing the second constraint.

Minimum-distance pair matching is a special case of a general class of problems known as minimum-cost network flow optimization (Rosenbaum, 1989). These problems, described with clarity by Bertsekas (1998), assume a directed graph consisting of a finite node set $\mathcal{N}$ and a finite edge set $\mathcal{E}$ with elements of the form (i, j) for $i, j \in \mathcal{N}, i \neq j$. A supply function $b : \mathcal{N} \longrightarrow \mathbb{Z}$ describes an integral number of units (of a commodity, perhaps) either supplied/produced at that node (if $b(n) > 0$) or demanded at that node (if $b(n) < 0$). In addition, a capacity function $d : \mathcal{E} \longrightarrow \mathbb{N}$ gives the maximum number of units that can be sent along each directed edge, and a cost function $c : \mathcal{E} \longrightarrow \mathbb{R}^+ \cup \{0\}$ gives the cost per unit transmitted over each edge. The optimization problem requires choosing a nonnegative integral amount of flow to be sent over each directed edge in such a way that flow of commodities is preserved at each node (i.e. the sum of incoming flow and supply at each node is equal to the sum of outgoing flow and absolute demand) and that edge capacities are respected, and to minimize the total cost paid among all such flows. Mathematically, the problem can be stated as follows:

$$\min \sum_{e \in \mathcal{E}} c(e) x_e \tag{1}$$

s.t.

$$\sum_{(k,n) \in \mathcal{E}, k \in \mathcal{N}} x_{(k,n)} + b(n) - \sum_{(n,k) \in \mathcal{E}, k \in \mathcal{N}} x_{(n,k)} = 0 \qquad \text{for all } n \in \mathcal{N} \tag{2}$$

$$0 \leq x_e \leq d(e) \qquad \text{for all } e \in \mathcal{E} \tag{3}$$

$$x_e \in \mathbb{Z} \qquad \text{for all } e \in \mathcal{E} \tag{4}$$

For convenience we will refer to the set of solutions $\mathbf{x} \in \mathbb{R}^{|\mathcal{E}|}$ satisfying both (2) and (3) as $\mathcal{F}$. A surprising and very useful aspect of minimum-cost network flow problems is that even if the integer constraint (4) is dropped, an integral solution is guaranteed to exist and can be computed in polynomial time (Bertsekas, 1991). This property, which comes from the totally unimodular property of network incidence matrices (Papadimitriou and Steiglitz, 1982, sec. 13.2), allows minimum-cost network flow problems to be solved efficiently at large scales even when closely related integer programs cannot be.

To see how matching fits into this framework, consider a bipartite graph connecting treated units $\mathcal{T} = \{\tau_1, \ldots, \tau_{N_1}\}$ to control units $\mathcal{C} = \{\kappa_1, \ldots, \kappa_{N_0}\}$; add to this graph an additional node ω we will call the sink, and edges connecting each control κ_j to the sink. The node set for the graph is $\mathcal{N} = \mathcal{T} \cup \mathcal{C} \cup \{\omega\}$, and the edge set consists of edges (τ_i, κ_j) for all $i \in \{1, \ldots, N_1\}$ and $j \in \{1, \ldots, N_0\}$, plus edges (κ_j, ω) for all $j \in \{, \ldots, N_0\}$. Define supply, capacity, and cost functions as follows, where δ_{ij} is some predefined multivariate distance denoting the similarity of treated unit i and control unit j:

$$b(n) = \begin{cases} 1 & \text{if } n = \tau_i \text{ for some } i \in \{1, \ldots, N_1\} \\ 0 & \text{if } n = \kappa_j \text{ for some } j \in \{1, \ldots, N_0\} \\ -N_1 & \text{if } n = \omega \end{cases}$$

$$d(e) = 1 \quad \text{for any } e \in \mathcal{E}$$

$$c(e) = \begin{cases} \delta_{ij} & \text{if } e = (\tau_i, \kappa_j) \text{ for some } i \in \{1, \ldots, N_1\}, j \in \{1, \ldots, N_0\} \\ 0 & \text{otherwise} \end{cases}$$

Essentially, we supply a unit of flow at each treated unit i and require it to be delivered to the sink, which requires the flow to be routed through some control unit j invoking a cost δ_{ij} on that treated-control edge. Since the control-sink edges are free of cost, the optimization problem reduces to finding the lowest total or average cost of all pairs. Notice that the upper capacity of 1 on each control-sink edge requires that each control be selected in at most one pair, enforcing matching without replacement. Formulating the problem in this way allows it to be solved quickly, even at large scales, by specialized software developed for minimum-cost network flow optimization (Bertsekas et al., 1994). The optimal matched design may be read off from the network flow solution by selecting pairs for which the flow variable of the corresponding edge has been set to one. For a more explicit description of the equivalence between these problems see Rosenbaum (1989).

While minimum-distance pair matching is commonly used in practice, it is frequently convenient to incorporate additional constraints or allow different matched set configurations. To preserve the fast computation associated with minimum-cost network flow optimization, it is generally necessary to formulate such modifications as network problems. As a simple example, fixed ratio matching (where $w \geq 1$ controls are matched to each treated unit) may be conducted by modifying the network just described to supply w units of flow at each treated unit and demand wN_1 at the sink. More involved modifications of the network problem to encapsulate more complex versions of the matching problem are described in Hansen and Klopfer (2006), Yang et al. (2012), and Pimentel et al. (2015a).

3.3 | Two optimization problems

We present two versions of the rolling enrollment matching optimization problem, each imposing a different restriction on the reuse of control subjects across distinct matched sets. Expanding the notation of Section 3.2, we use δ_{ijk} to represent a predetermined covariate distance between treated subject i over a fixed-length pre-exposure period and the kth version of control subject j over the associated fixed-length pre-exposure period. We also let m_j be the number of different versions of control subject j we consider matching to (i.e. the number of different possible longitudinal alignments).

Problem A (1:w matching without any reuse of controls)

$$
\begin{aligned}
&\min \sum_{i=1}^{N_1} \sum_{j=1}^{N_0} \sum_{k=1}^{m_j} x_{ijk} \delta_{ijk} \\
&s.t. \\
&\sum_{j=1}^{N_0} \sum_{k=1}^{m_j} x_{ijk} = w \qquad \textit{for all } i \in \{1, \ldots, N_1\} \\
&\sum_{i=1}^{N_1} \sum_{k=1}^{m_j} x_{ijk} \leq 1 \qquad \textit{for all } j \in \{1, \ldots, N_0\} \\
&x_{ijk} \in \{0, 1\} \qquad \textit{for all } i \in \{1, \ldots, N_1\}, j \in \{1, \ldots, N_0\}, k \in \{1, \ldots, m_j\}
\end{aligned}
$$

In this problem we require that if some treated unit is matched to some version of a control subject, no other treated unit may be matched to any other version of that control subject. This ensures that when we have N_1 treated subjects, we include wN_1 distinct control subjects in our final design.

Problem B (1:w matching with reuse of non-identical versions)

$$
\begin{aligned}
&\min \sum_{i=1}^{N_1} \sum_{j=1}^{N_0} \sum_{k=1}^{m_j} x_{ijk} \delta_{ijk} \\
&s.t. \\
&\sum_{j=1}^{N_0} \sum_{k=1}^{m_j} x_{ijk} = w \qquad \textit{for all } i \in \{1, \ldots, N_1\} \\
&\sum_{k=1}^{m_j} x_{ijk} = 1 \qquad \textit{for all } i \in \{1, \ldots, N_1\}, j \in \{1, \ldots, N_0\} \\
&\sum_{i=1}^{N_1} x_{ijk} = 1 \qquad \textit{for all } j \in \{1, \ldots, N_0\}, k \in \{1, \ldots, m_j\} \\
&x_{ijk} \in \{0, 1\} \qquad \textit{for all } i \in \{1, \ldots, N_1\}, j \in \{1, \ldots, N_0\}, k \in \{1, \ldots, m_j\}
\end{aligned}
$$

In this problem two distinct versions of the same control subject may be matched to two different treated units. However, no individual matched set may contain multiple versions of the same control subject, and each specific version of a control subject may be used only once in the match as a whole.

3.4 | Network flow algorithms

We now present network flow algorithms that solve Problems A and B. To begin, consider the network structure described in Section 3.2, a directed bipartite graph from treated subjects to control subjects augmented with a sink node to which each control connects. We proceed by expanding this structure to incorporate the additional constraints in Problems A and B.

To solve Problem A, we use a graph with all the same nodes as the one in Section 3.2 (a node τ_i for each treated unit $1, \ldots, N_1$, a node κ_j for each distinct control subject $1, \ldots, N_0$ and a sink node ω) but add some additional nodes.

We also include a node κ_{jk} ($j \in \{1, \ldots, N_0\}$ and $k \in \{1, \ldots, m_j\}$) for each distinct version of a control. Formally, we define the node set as

$$\begin{aligned}\mathcal{N}_A &= \left\{\tau_1, \ldots, \tau_{N_1},\ \kappa_1, \ldots, \kappa_{N_0},\ \omega,\ \kappa_{11}, \ldots, \kappa_{1m_1}, \kappa_{21}, \ldots, \kappa_{N_0 m_{N_0}}\right\} \\ &= \mathcal{N} \ \cup \{\kappa_{jk}\}_{j=1,\ldots,N_0,\ k=1,\ldots,m_j}.\end{aligned}$$

Here $\mathcal{N}$ represents the node set from Section 3.2. We specify the edge set as follows:

$$\begin{aligned}\mathcal{E}_A = &\left\{(\tau_i, \kappa_{jk}) : i \in \{1, \ldots, N_1\}, j \in \{1, \ldots, N_0\}, k \in \{1, \ldots, m_j\}\right\} \\ &\cup \left\{(\kappa_{jk}, \kappa_j) : j \in \{1, \ldots, N_0\}, k \in \{1, \ldots, m_j\}\right\} \\ &\cup \left\{(\kappa_j, \omega) : j \in \{1, \ldots, N_0\}\right\}.\end{aligned}$$

To complete the specification of the network flow problem allowing fixed-ratio 1:w matching, we set the supply, capacity, and cost functions as follows.

$$\begin{aligned}b_A(n) &= \begin{cases} w & \text{if } n = \tau_i \text{ for some } i \in \{1, \ldots, N_1\} \\ -wN_1 & \text{if } n = \omega \\ 0 & \text{otherwise} \end{cases} \\ d_A(e) &= 1 \quad \text{for any } e \in \mathcal{E}_A \\ c_A(e) &= \begin{cases} \delta_{ijk} & \text{if } e = (\tau_i, \kappa_{jk}) \text{ for some } i \in \{1, \ldots, N_1\}, j \in \{1, \ldots, N_0\}, k \in \{1, \ldots, m_j\} \\ 0 & \text{otherwise.} \end{cases}\end{aligned}$$

In simple terms, the individual-version control nodes κ_{jk} form an additional “layer" in the graph, lying between the treated layer of nodes τ_i and the layer of distinct control subjects κ_j. Covariate distances are assigned to the edges connecting the treated layer and the individual-version layer (since covariate distances are a function of the fixed-length pre-exposure period which changes with the alignment chosen for an individual control subject). However, all flow through individual-version control nodes κ_{jk} for any j is then pooled at a single node κ_j, and since the outflow from this node to ω must not exceed one, no more than one version of any individual control subject may be selected into the match.

To solve Problem B, we construct a node set N_B with a node τ_i for each of N_1 treated units, a node κ_{jk} for each of the $\sum_{j=1}^{N_0} m_j$ versions of individual control subjects, and a sink node ω; in addition, we include a node ζ_{ij} for each distinct combination of treated subject and distinct control subject (meaning there are $N_1 N_0$ of these nodes in total). We define the edge set and the supply, capacity, and cost functions as follows.

$$\begin{aligned}\mathcal{E}_B = &\left\{(\tau_i, \zeta_{ij}) : i \in \{1, \ldots, N_1\}, j \in \{1, \ldots, N_0\}\right\} \\ &\cup \left\{(\zeta_{ij}, \kappa_{jk}) : i \in \{1, \ldots, N_1\}, j \in \{1, \ldots, N_0\}, k \in \{1, \ldots, m_j\}\right\} \\ &\cup \left\{(\kappa_{jk}, \omega) : j \in \{1, \ldots, N_0\}, k \in \{1, \ldots, m_j\}\right\}\end{aligned}$$

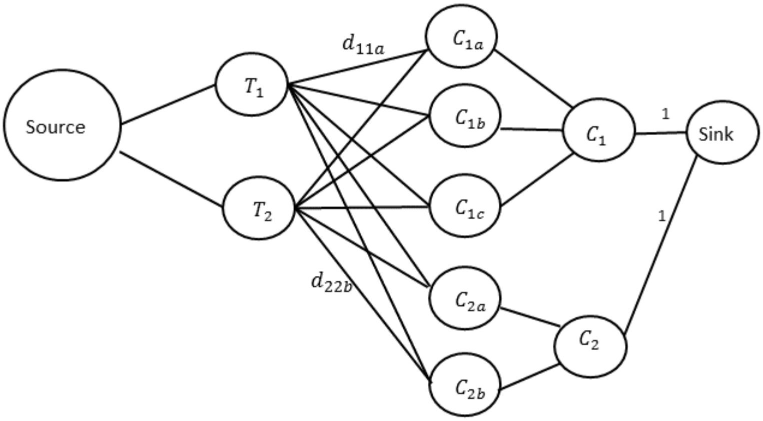

FIGURE 2 Network formulation for Problem A. In this figure, each circle is a node representing a treated subject, control version, or unique control subject, progressing in that order from left to right. Each line is an edge. Edges joining treated nodes such as τ_1 and control version nodes such as κ_{11} represent potential matches and are weighted with a cost corresponding to the δ_{ijk} of the formulation above. Flow passes from the treated nodes on the left to the sink on the right, where capacity constraints of 1 on the edges that connect unique comparison nodes such as κ_1 to the sink ensure that no more than one version per unique control is matched.

$$b_B(n) = \begin{cases} w & \text{if } n = \tau_i \text{ for some } i \in \{1, \dots, N_1\} \\ -wN_1 & \text{if } n = \omega \\ 0 & \text{otherwise} \end{cases}$$

$$d_B(e) = 1 \quad \text{for any } e \in \mathcal{E}_B$$

$$c_B(e) = \begin{cases} \delta_{ijk} & \text{if } e = (\zeta_{ij}, \kappa_{jk}) \text{ for some } i \in \{1, \dots, N_1\}, j \in \{1, \dots, N_0\}, k \in \{1, \dots, m_j\} \\ 0 & \text{otherwise} \end{cases}$$

The networks for both Problems A and B contain three layers plus a sink, but instead of having a layer for distinct treated subjects followed by a layer for versions of controls followed by a layer for distinct control subjects (as did the network for Setting A), this network follows the treated layer with a layer for all distinct treated-control pairings, then includes a layer for versions of controls that in turns connects directly to the sink. The edge costs are assigned to the edges between the distinct-pairing layer and the distinct-version layer. In contrast to Setting A, this network allows distinct treated units to match to copies of the same control, since the distinct-control layer permitting only one version per subject to be included has been removed. However, the distinct-pairing layer enforces the constraint that no more than one version of a control be matched to the same treated unit.

3.5 | Correspondence between reduced-form and network formulations

In Section 3.3 we presented two matching problems we hope to solve, Problems A and B, and in Section 3.4, we presented network flow problems designed to solve them, denoted Network A and Network B . In the current section, we formally demonstrate that Network A correctly solves Problem A by presenting an objective-cost-preserving bijective mapping between solutions to the original problem and solutions to the network formulation. The correspondence

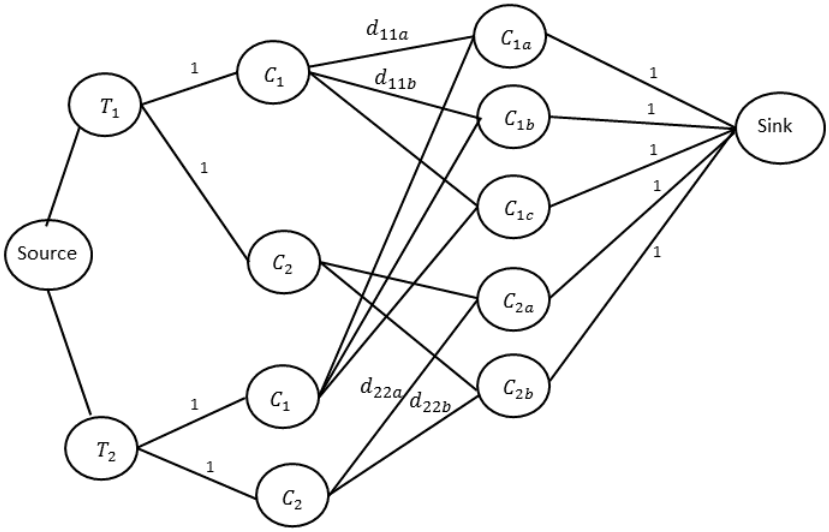

FIGURE 3 Network formulation for Problem B. This figure adopts a similar structure to that for Problem A but exchanges the columns of nodes representing control versions and unique controls. In addition, it includes a node for every combination of treated subject and unique control subject (ζ_{ij}). Duplicating the unique control subject nodes allows different treated subjects to match to different versions of the same control. Edges connecting these classes of nodes have a capacity of one, ensuring that no treated subject matches to more than one version of a given control subject. Likewise, the edges connecting nodes representing control versions have a capacity of one so that no version matches to more than one treated subject. The costs δ_{ijk} of pairing a treated individual i to a version of a control subject jk are borne by the $(\zeta_{ij}, \kappa_{jk})$ edges.

between Network B and Problem B may be proved by almost identical arguments.

Let $\mathcal{F}_A^P \in \mathbb{Z}^{N_1 N_0}$ and $\mathcal{F}_A^N \in \mathbb{Z}^{|\epsilon_A|}$ be the feasible sets of Problem A and Network A respectively. Define a mapping $F : \mathcal{F}_A^P \longrightarrow \mathbb{Z}^{|\epsilon_A|}$ where $\mathbf{x}' = F(\mathbf{x})$ is constructed as follows :

$$\begin{aligned} x'_{(\tau_i,\kappa_{jk})} &= x_{ijk} \\ x'_{(\kappa_{jk},\kappa_j)} &= \sum_{i=1}^{N_1} x_{ijk} \\ x'_{(\kappa_j,\omega)} &= \sum_{i=1}^{N_1}\sum_{k=1}^{m_j} x_{ijk} \end{aligned}$$

To establish the result, we rely on the following two lemmas, the proofs of which are deferred to the appendix.

Lemma 1 *$F(\mathbf{x}) \in \mathcal{F}_A^N$ for any $\mathbf{x} \in \mathcal{F}_A^P$.*

Lemma 2 *F is a bijection between $\mathcal{F}_A^P$ and $\mathcal{F}_A^N$, with inverse $F^{-1} = G$ defined as follows:*

$$G(\mathbf{x}') = \mathbf{x} \quad \textit{such that} \quad x_{ijk} = x_{(\tau_i,\kappa_{jk})} \forall i, j, k.$$

Theorem 2 *A solution $\mathbf{x}$ is optimal for Problem A if and only if $\mathbf{x} = F^{-1}(\mathbf{x}')$ where $\mathbf{x}'$ is optimal for Network A.*

Proof The objective cost $C_{\mathbf{x}_0}^P$ of any feasible solution $\mathbf{x}_0$ for Problem A, $\sum_{i=1}^{N_1}\sum_{j=1}^{N_0}\sum_{k=1}^{m_j} x_{ijk}$, is identical to the objective cost $C_{F^{-1}(\mathbf{x}_0)}^N$ of $F^{-1}(\mathbf{x}_0)$ in Network A by Lemma 2. Since $C_{\mathbf{x}}^P \leq C_{\mathbf{x}_0}^P$ for any $\mathbf{x}_0 \in \mathcal{F}_A^P$, $C_{F^{-1}(\mathbf{x})}^N \leq C_{F^{-1}(\mathbf{x}_0)}^N$ for any $\mathbf{x}_0 \in \mathcal{F}_A^N$. The reverse direction holds for any $\mathbf{x}'$ optimal for Network A by an identical argument.

4 | SIMULATION STUDY

The arguments presented thus far articulate the conceptual advantage of allowing for the TALI assumption and point to the optimality of our GroupMatch implementation. To lend these claims empirical support, we conduct a simulation study in which we apply each of several matching techniques, representing the industry standard for health policy evaluations with rolling enrollment, to a simulated data set. We wish to learn which features of the simulated data and matching process are associated with different aspects of simulation performance, so in simulating the data and performing the matching, we vary three parameters (Table 1). Together, these parameters form twelve unique combinations, and for each of these combinations we conduct a simulation of 1,000 iterations.

Parameter	Value(s)
Number of versions per potential comparison	{4, 12}
Matching ratio	{1T:3C, 1T:4C, 1T:5C}, fixed
Functional form of outcome	{Linear, Nonlinear}

TABLE 1 Parameter values tested in our simulation study.

4.1 | Design

4.1.1 | Data Generation

We simulate data sets to have the properties common in rolling enrollment matching scenarios, particularly for health policy programs. Each simulation consists of 1,000 total individuals. The proportion of these individuals whom we assign to the treated group is contingent on the matching ratio. For 3:1 matching, we assign 1/5 subjects to treatment, for 4:1 matching we assign 1/6, and for 5:1 matching we assign 1/7. These choices were made to ensure enough potential comparisons to solve each matching problem (both with and without replacement), with some comparison individuals remaining unmatched after matching.

For each unique individual we randomly generate four baseline covariates, intended to represent demographic characteristics such as age, race, or gender that remain constant across observations of the same individual. We then generate either 4 or 12 "versions" of each potential control individual by assigning several different pseudo-enrollment periods to each potential control.

Because an acute event, such as a stroke or heart attack, often serves as a "trigger event" for a rolling enrollment intervention, we randomly assign an acute event indicator to each potential control version, at a random time; all treatment cases are assumed to have an acute event. Conditional on the acute event indicator, we generate four time-varying covariates intended to represent prior utilization measures, such as Medicare spending or hospitalizations, that reflect the case's health status at that point in time. For example, cases – treatment or potential control – that have an acute event are more likely to have higher expenditures and hospitalization rates. We match on all eight covariates, the four baseline variables and the four time-varying variables. Finally, we generate a continuous outcome variable in two ways. First, based on a linear model that conditions on all of these background characteristics, with covariate coefficients ranging from $\log(1.25)$ to $\log(10)$, a treatment coefficient of 1, and an intercept of 0. Second, we generate the outcome variable as a function of all eight background characteristics, plus an interaction between the background characteristics and treatment, scaled so that the treatment effect remains 1.

4.1.2 | Matching

We apply six different matching techniques to each randomly generated data set. The basis for each matching approach is a set of fitted propensity scores estimated via logistic regression. Unless specified otherwise, optimal matching as implemented in R's optmatch package (Hansen, 2007) forms the basis of all matching approaches.

- **Exact matching.** This technique uses exact-matching constraints on the observation time period to avoid assigning more than one duplicate of the same comparison to a given treatment subject.
- **Rolling entry matching (REM).** This technique was developed for rolling enrollment matching problems (Witman et al., 2019), seeking both to match no more than one version of the same unique control to a given treatment subject and to retain no more than one version of a given unique control in the final matched sample. We test this approach both with and without replacement using version 2.0.1 of the R package rollmatch (Chew et al., 2018), which implements greedy rather than optimal matching.
- **Sequential matching.** With sequential matching, we assign w controls to each treatment case by performing matching in w rounds. In a first round, with all potential control versions in the pool, we assign one control to each treatment case. We then remove any version of any control that has been matched from the pool before proceeding to the next round of matching. We repeat this process until we have assigned w unique controls to each treatment case.

- **GroupMatch.** GroupMatch is the novel approach proposed in this paper. We implement it both allowing and forbidding replacement, where "replacement" means that different treatment subjects can match to different versions of the same unique control, but no treatment can match to more than one version of a give control, and no two treatment subjects can match to the same version of a control. "Without replacement" and "with replacement" GroupMatch options refer to Problems A and B as formulated in Section 3.3. We use a development version of our GroupMatch R package.

After performing matching using each of these techniques, we extract the matched sets and use them to assess the quality of the resulting matched sample.

4.2 | Results

The simulation results indicate that GroupMatch performs very strongly, in absolute terms and relative to the other methods tested. We examined performance on both covariate balance and on mean squared error, estimating the treatment effect using a simple difference in means in each simulation data set. Throughout, conditions where the treatment effect is linear in the covariates produced almost indistinguishable results to conditions where the treatment effect was nonlinear in the covariates, so we present only the former.

4.2.1 | Balance and Treatment Effect Estimation

To assess covariate balance, we calculate the standardized difference on each matching variable between the treatment groups in the matched sample, defined as the difference in means for that variable divided by the standard deviation in the treatment group (Rosenbaum and Rubin, 1985; Rosenbaum, 2010, sec. 9.1). The top row of Figure 4 plots the mean absolute standardized differences for three sets of variables: (1) the baseline (time-static) predictors, (2) the acute event indicator, and (3) time-varying predictors assumed to be affected by the acute event. We find that the number of control versions only has minimal effect on the results, so we present results only for the 12-version scenario for simplicity. As the figure shows, across each of the scenarios (matching ratios and predictor types), GroupMatch with replacement achieves consistently excellent balance. Balance tends to be worse for GroupMatch without replacement than for other approaches, except for Rolling Enrollment Matching without replacement; we believe the poor performance of GroupMatch without replacement, and likely also of Rolling Enrollment Matching without replacement, stems from the comparatively small pool of unique potential comparison units. In cases with vastly more unique potential comparison units than treated units, as is the norm in many real-world applications, we would expect matching with and without replacement to perform comparably for both approaches.

In addition to examining the covariate balance, we also assess the ability of each method to estimate the true treatment effect, calculating the mean squared error of each estimation procedure. These values are plotted in the bottom row of Figure 4. GroupMatch approaches perform the best of all alternatives, with minimal MSE regardless of the matching ratio. Intriguingly, GroupMatch without replacement performs quite well despite obtaining worse than average balance.

4.2.2 | Run Time

Although unrelated to its efficacy in selecting high-quality comparison groups, run time is an important consideration when performing matching with real-world data. As part of the simulation study, we also timed the matching step for

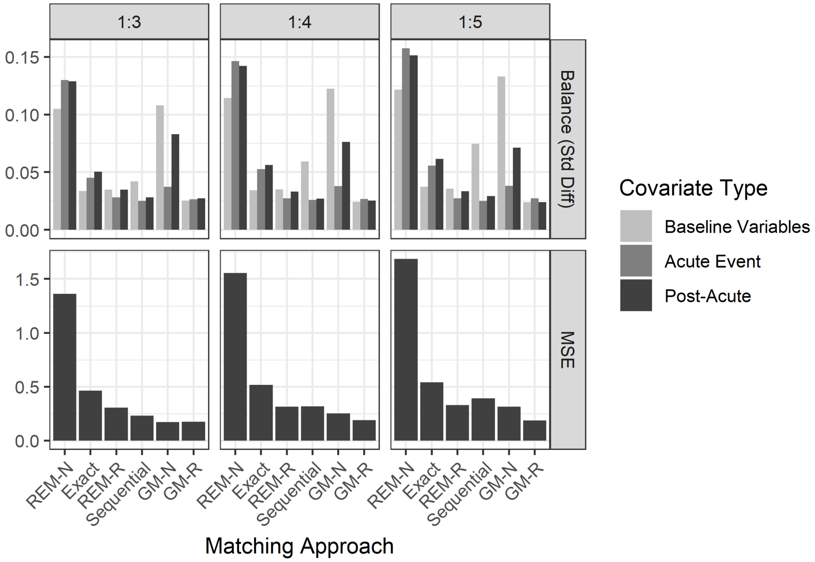

FIGURE 4 This figure summarizes the results of the simulation study. Each column represents a different matching ratio, ranging from 1:3 on the left to 1:5 on the right. Each row represents a different diagnostic, with balance assessments on the top, measured in absolute standardized differences, and mean squared error on the bottom. In each panel, we compare among six different matching approaches: REM without replacement (REM-N), exact matching on enrollment/pseudo-enrollment period (Exact), REM with replacement (REM-R), sequential matching (Sequential), GroupMatch without replacement (GM-N), and GroupMatch with replacement (GM-R).

each approach and summarized the run times across iterations.

Matching Approach	Number of Versions	Matching Ratio		
		1:3	1:4	1:5
Exact	4	4	4	4
	12	36	38	39
GM-N	4	14	17	17
	12	12	10	11
GM-R	4	19	19	18
	12	32	28	26
Sequential	4	11	12	13
	12	30	35	40
REM-N	4	2	2	2
	12	2	2	2
REM-R	4	1	1	1
	12	1	2	2

TABLE 2 Run time, in seconds, of each matching algorithm, averaged across the simulation iterations. Rows refer to combinations of matching approach and number of versions per unique control (4 or 12). Columns identify the matching ratio.

The three approaches that achieved good balance and mean squared error in the simulation study – GroupMatch with replacement, exact matching, and sequential matching – are slower than the less-successful approaches, but fairly comparable to one another. We also see that run times typically increase as the number of versions of each comparison increases, though the number of versions does not appreciably affect GroupMatch without replacement (GM-N) or the two REM approaches. In real-world applications, we have found that exact-match variables and other calipers can dramatically improve the run time for GroupMatch; across a set of problems with hundreds of thousands of observations, run times range from 15 minutes to four hours, depending on the number of treated observations. These run times are typical, in our experience, for full optimal matching, so we believe that the more complex network configuration does not lead to noticeably slower processing.

5 | CASE STUDY: IMPACT OF FALLS

As a case study, we re-analyze a data set used to estimate the impact of falls on the Medicare Advantage population enrolled in the Group Health Western Washington Integrated Group Practice (IGP), located near Seattle, Washington, USA (Bohl et al., 2010). The data set tracks the post-fall total cost of care for Medicare Advantage enrollees who fell. To estimate the effect of falls on total cost of care, we must compare these fallers to otherwise-similar non-fallers.

In the original study, the authors assigned a single pseudo-fall date to each potential control before selecting a control group using frequency matching on age and gender. With our new GroupMatch approach, we expand the control pool to include nine pseudo-fall dates per potential control. Reorganizing the data in this way requires us to use the potential control observations' post-intervention time points in matching, preventing us from using the matched sample to estimate program impacts.

5.1 | Data

The data comprise eligible Medicare beneficiaries from Group Health's Western Washington IGP who were alive and age 67 or older on January 1, 2004. The study period ranged from January 1, 2004 and December 31, 2006; to be eligible for the study, beneficiaries must have been enrolled in Medicare for at least two years between January 1, 2002 and December 31, 2006. We excluded beneficiaries with any record of a fall-related episode between January 1, 2002 and December 31, 2003.

From this data set, we constructed the potential control pool by assigning nine quarterly pseudo-index dates to each non-faller. The non-faller must be observed in the data for one year before each pseudo-fall date to ensure adequate data for matching. The final sample contained 3,517 unique fallers (treatment group) and 8,956 unique non-fallers (potential control group).

Group Health's data warehouse contains rich and varied information about each Medicare Advantage enrollee, including time-varying medical utilization information – for example, Medicare expenditures – as well as information on chronic conditions and medical risk. We use these variables to develop our matching approach.

5.2 | Methods

We compare several methods for comparison group selection: REM, sequential matching, and GroupMatch, as previously defined. In each case, the propensity score model includes the beneficiary's age, sex, measures of baseline health care spending and utilization, and chronic conditions. We also compare these matching approaches to the balance obtained in the original study, which randomly assigned a single pseudo-fall date to each potential control subject and frequency-matched on age and gender only. This set of approaches allows us both to investigate the gains associated with the TALI assumption and to compare performance among rolling enrollment matching methods.

Without information about the true treatment effect, we must compare the candidate approaches on balance alone. For simplicity, we show the distribution of absolute standardized differences obtained under each approach for all variables included in the propensity score model. We depict those results in Figure 4.

As we see in the figure, GroupMatch with replacement and the sequential matching approach perform very comparably, with similarly tight distributions and no standardized differences greater than 0.1 in absolute value. GroupMatch without replacement and the two REM approaches perform comparably, though GroupMatch without replacement achieves much more consistent balance across all the matching variables, as the narrower distribution indicates.

In a data set of this size, GroupMatch without replacement is a sub-optimal approach because the ratio of unique potential control subjects to unique treatment subjects barely exceeds two. We believe that with a more generous ratio of unique potential control subjects to unique treatment subjects, GroupMatch without replacement would perform better.

Of the alternative methods considered, sequential matching performs quite well in both the simulation and empirical settings, with comparable balance to GroupMatch in the application and only slightly worse MSE in the simulation. Although the two approaches appear evenly matched, our experience using sequential matching to solve real health policy problems leads us to prefer GroupMatch. Compared to GroupMatch, sequential matching is extremely laborious to implement and time-consuming to run, especially in larger datasets than considered here. In fact, the challenges we faced implementing sequential matching to solve large-scale problems inspired us to search for alternatives, a

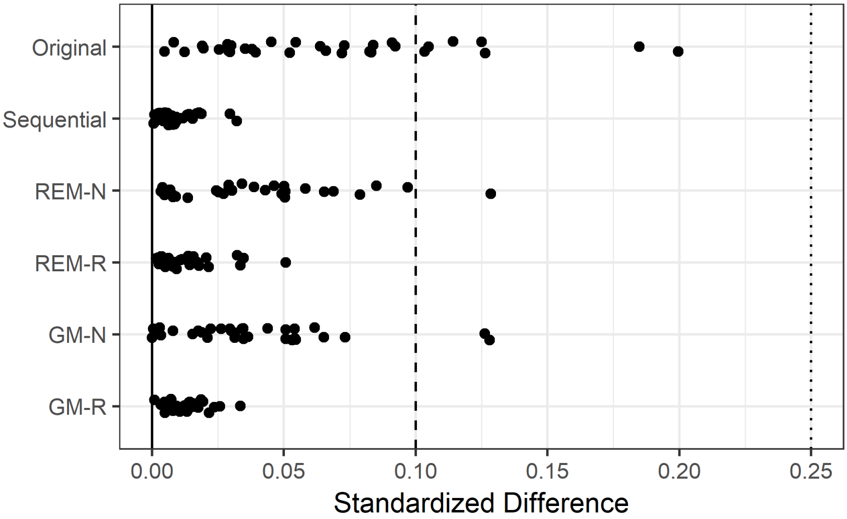

FIGURE 5 Distribution of absolute standardized differences by approach, where Original is the original data set with only one version of each unique potential comparison; Sequential is sequential matching; REM-N is REM without replacement; REM-R is REM with replacement; GM-N is GroupMatch without replacement; and GM-R is GroupMatch with replacement.

search that led us to develop GroupMatch. Thus, we prefer GroupMatch for its simplicity, ease of implementation, and efficiency.

6 | APPROACHES TO INFERENCE

While our chief aim is to propose a comparison group selection strategy, the GroupMatch framework, we must also consider how to obtain inferences for the resulting causal effect estimates. One option is to consider an asymptotic regime in which datasets used for matching grow larger and larger but the number of units in any matched set remains fixed, and to conduct inference based on a resulting central limit theorem. When the conditions assumed in Theorem 1 hold and matches are produced by Problem A, so that all units involved in the match come from distinct individuals, such large-sample inference is possible using the methods of Abadie and Imbens (2012) with relatively minor modifications. In particular, following these authors, we may conduct asymptotic inference on the population ATT using a normal approximation with the following variance estimator (in the notation of Section 2).

$$\widehat{\text{Var}}\left(\widehat{\Delta}\right) = \frac{1}{N_1 - 1} \sum_{i=1}^{n} \sum_{t=1}^{T} 1\left\{Z_i^t = P\right\} \left(Y_i^t - \frac{1}{w} \sum_{j=1}^{n} \sum_{t'=1}^{T} M_{it,jt'} Y_j^{t'} - \widehat{\Delta} \right)^2$$

This estimator is simply the sample variance of the matched pair differences in the final design, with an $N_1 - 1$ bias correction. For a more detailed description of how the Abadie and Imbens (2012) framework applies to this setting, see the online supplement.

A second approach to developing inference that does not rely on asymptotics is to consider an ideal randomized experimental design that GroupMatch designs approximate, and to base inference for GroupMatch on the corresponding inference methods for the randomized design. This idea is part of a larger tradition of presenting matched designs as approximations to randomized trials that would have been conducted had resources and practical constraints allowed (Rubin, 2007; Rosenbaum, 2010). In particular, many matched designs are analyzed as though pairs or matched sets had been formed prior to treatment, with treatment subsequently allocated via complete randomization within the pairs or sets.

GroupMatch designs do not fit comfortably into this tradition of analogous randomized experiments, since pairs include individuals at disparate points in time and could not reasonably have been formed in advance of all treatment assignments. In addition, matches produced by solving Problem B may include multiple versions of the same individual across individual matched pairs or sets. If we restrict attention to matches produced by Problem A, however, a similar analogy can be developed for a simple paired sequential experiment, which we now describe.

Suppose patients enroll in our study sequentially, each patient being immediately assigned either to treatment or to control, and either paired to a previously enrolled patient or assigned to the as-yet-unpaired patient group. At any given timepoint, there will be some combination of paired patients and unpaired patients enrolled. When a new patient enters we make assignments as follows. We measure the patient's baseline covariates (possibly including lagged covariate values) and assess his/her similarity to all other unpaired patients on the same covariates. We then make a list of all unpaired patients previously enrolled who are sufficiently similar on baseline covariates (with respect to some predefined criterion). Suppose there are k such patients, where k can be equal to zero. We then choose uniformly at random among $k+1$ possibilities: pairing the new patient to each of the k patients already enrolled, or enrolling this patient as an unpaired individual. If paired, the patient will be assigned to the opposite treatment condition to its paired counterpart; if left unpaired, it is assigned to treatment or control with equal probability. Imagine running such an experiment until a large number of pairs has been collected and focusing on the matched pairs for analysis. Distinct individuals' marginal probabilities of assignment to treatment or control are identical, so that paired individuals are equally likely to receive both treatment and control. Similar experiments have been described and analyzed mathematically by Kapelner and Krieger (2014). Note that the design can easily be generalized to settings with matched sets containing $k > 1$ controls for each treated unit.

The connection to GroupMatch designs is straightforward. The final design produced by GroupMatch (when run without duplication across matched sets) contains matched pairs or sets of a given size, each with exactly one unit receiving treatment. Just as in the sequential experiment, these units may be aligned to different timepoints, but they share similar lagged covariate values. Under time-agnostic L-ignorability and covariate L-exogeneity, differences in means within these sets should be unbiased for treatment effects, just as treated-control differences in the randomized trial are unbiased for treatment effects.

To develop finite-sample inference, we impose a simple model on the matched pairs produced in which matched sets are independent of one another and all allocations of treatment with exactly one treated unit per set are equally likely. Assume a sharp null hypothesis $Y_i^T(P) = Y_i^t(0)$ for all i, t; this corresponds to a setting with no effect of treatment whatsoever after follow-up time P. Under this null hypothesis, the observed outcomes in our matched sets would have remained identical had treatment been allocated differently among the matched patients. Thus, we may define a test statistic T as a function of the observed treatment and outcomes in our data, and under the null we can also compute its null distribution by considering alternate permutations of treatment. This framework, which has been developed extensively by Rosenbaum (2002), can be extended to construct confidence intervals for finite-sample treatment effects of various kinds and to relax the assumption of equal-probability of treatment via sensitivity analyses. Under the proposed model for randomization inference, these methods can be applied directly to matched sets created using GroupMatch.

7 | DISCUSSION

As formulated, the TALI assumption usefully expands the comparison pool beyond the limiting assumptions inherent in assigning each potential comparison an arbitrary pseudo-enrollment date, and the simulation and case studies provide empirical support for its efficacy. Despite these advantages, the inferential framework we propose in this paper is far from iron-clad, and here we elaborate on some of its limitations.

In practice, time point-agnosticism may not always be a reasonable assumption; if for example the database covers a very long time course then it may not be wise to compare a control subject from a very early point in the time scale to a treated subject from the very end of the time scale. This situation often rises in long-term health policy evaluations and health programs which target mild chronic persons. In such settings, however, time point agnosticism may still hold approximately within shorter time windows, so that matched pairs may provide reliable estimates so long as they are not too distant in time. The conceptual approach, and the GroupMatch implementation, are flexible enough to accommodate these restrictions, which are very easily implemented through the network methods of section 3. For example, let C be the maximum separation in time between two paired subjects' time-alignments. All edges (τ_i, κ_{jk}) in Network A and edges (η_{ij}, κ_{jk}) connecting a treated subject i with a control version jk more than C timepoints apart may simply be deleted from the graph. This is an example of caliper matching as described by Hansen and Klopfer (2006). Removing edges from the network problem in this way has computational benefits in addition to any statistical ones, since it increases the network's sparsity (Pimentel et al., 2015a).

Another issue is that the methods for inference described in Section 6 are appropriate for matches produced by solving Problem A, but not for those produced by solving Problem B. In particular, the reuse of different versions of the same control individual across matched sets in solutions to Problem B breaks the analogy with the sequential paired random experiment of Section 6. Furthermore, while standard variance estimators for matching with replacement (Abadie and Imbens, 2006; Hill and Reiter, 2006) exist for simpler settings, these methods do not apply immediately to solutions to Problem B. The situation in Problem B is an intermediate case between matching without replacement and matching with replacement; although exact versions of control subjects are not repeated from set to set, two different time-alignments of the same control unit may appear in different sets, and they may have highly correlated outcomes. Simple multiplication of matching weights, an important component of many variance estimation schemes for matching with replacement, is not appropriate here, since different versions of the same individuals may have different outcome values. However, treating the different versions as independent observations without accounting for intra-subject correlation will lead to anticonservative inference. Further work is needed to formulate compelling methods of variance estimation and inference for this case.

While in the current work we focus on a setting where subjects receiving treatment at some point are sampled separately from subjects who remain always in the control condition, the same statistical framework and matching algorithms might also be used to compare treated subjects after treatment enrollment to earlier versions of treated subjects before enrollment. While this design has disadvantages, including smaller effective sample sizes (especially in cases where “true" controls never receiving treatment vastly outnumber the treated group), it may allow for more plausible ignorability assumptions since all subjects are now sampled from the same population. Such a before-after treatment match could be conducted alongside a treated-control match as described above to provide a second estimate of the causal effect and to test for unmeasured bias as described in Rosenbaum (2002, ch. 8), or perhaps the designs could be combined in an approach similar to difference-in-differences.

Future work may also expand the range of settings to which we apply the proposed methods. Consider a longitudinal setting where different treated subjects have baseline histories of differing lengths prior to treatment. This situation is common in Medicaid studies, where beneficiaries who gain Medicaid eligibility recently have limited baseline histories. As an example, consider a scenario in which some beneficiaries have six months of baseline data, but others only have one month. Rather than excluding individuals who do not have a fully observed baseline history, we can define separate propensity score models for each available length of baseline data (a one-month propensity score and a six-month propensity score). Treated units would enter into the matching problem according to the length of their baseline history, but we would create multiple versions of each comparison unit: one based on the one-month propensity score, and one based on the six-month propensity score, if that data is available. We can then apply Problem A to select comparison subjects for treated subjects, imposing an exact-match restriction to permit only matches between versions with identical length of baseline data and defining distances based on the propensity score model corresponding to that baseline period. In this way, a single comparison subject could be matched to a treated subject with either a six-month baseline history or a one-month baseline history, but not both, and only according to the equivalently defined propensity score.

acknowledgements

The authors gratefully acknowledge the invaluable assistance of Alex Bohl, who provided the falls case study data. The authors also thank Aden Forrow and Bonnie Harvey for useful suggestions and design consultations in the project's formative stages, and Jas Sekhon for additional helpful comments.

appendix

Proof of Theorem 1. If $Z_i^t = P$, then $M_{it,jt'} = 1$ only if $Z_j^{t'} = 0$. By consistency of potential outcomes, we can write the following:

$$\begin{aligned}
\widehat{\Delta} &= \frac{1}{N_1}\sum_{i=1}^{n}\sum_{t=1}^{T} 1\left\{Z_i^t = P\right\}\left[Y_i^t(P) - \frac{1}{w}\sum_{j=1}^{N}\sum_{t'=1}^{T} 1\left\{Z_j^{t'} = 0\right\} M_{it,jt'} Y_j^{t'}(0)\right] \\
&= \frac{1}{N_1}\sum_{i=1}^{n}\sum_{t=1}^{T} 1\left\{Z_i^t = P\right\}\left[Y_i^t(P) - Y_i^t(0)\right] + \frac{1}{N_1}\sum_{i=1}^{n}\sum_{it=1}^{T} 1\left\{Z_i^t = P\right\}\left[Y_i^t(0) - \frac{1}{w}\sum_{j=1}^{N}\sum_{t'=1}^{T} 1\left\{Z_j^{t'} = 0\right\} M_{it,jt'} Y_j^{t'}(0)\right] \\
&= \Delta + \frac{1}{N_1}\sum_{i=1}^{n}\sum_{t=1}^{T} 1\left\{Z_i^t = P\right\}\left[\mu_0(X_i^{t-P}, \ldots, X_i^{t-P-L}) - \frac{1}{w}\sum_{j=1}^{N}\sum_{t'=1}^{T} 1\left\{Z_j^{t'} = 0\right\} M_{it,jt'}\mu_0(X_j^{t'-P}, \ldots, X_j^{t'-P-L})\right] \\
&+ \frac{1}{N_1}\sum_{i=1}^{n}\sum_{t=1}^{T} 1\left\{Z_i^t = P\right\}\left[\left(Y_i^t(0) - \mu_0(X_i^{t-P}, \ldots, X_i^{t-P-L})\right) - \frac{1}{w}\sum_{j=1}^{N}\sum_{t'=1}^{T} 1\left\{Z_j^{t'} = 0\right\} M_{it,jt'}\left(Y_j^{t'}(0) - \mu_0(X_j^{t'-P}, \ldots, X_j^{t'-P-L})\right)\right]
\end{aligned}$$

By our assumptions of exact matching and by timepoint agnosticism (implicit in the μ_0 notation), the second additive term is equal to zero since covariate histories of matched units are identical. We rewrite the remaining terms representing the disturbance terms $Y_i^t(0) - \mu_0(X_i^{t-P}, \ldots, X_i^{t-P-L})$ by ξ_i^t.

$$\widehat{\Delta} = \Delta + \frac{1}{N_1}\sum_{i=1}^{n}\sum_{t=1}^{T} 1\left\{Z_i^t = P\right\}\left[\xi_i^t - \frac{1}{w}\sum_{j=1}^{N}\sum_{t'=1}^{T} 1\left\{Z_j^{t'} = 0\right\} M_{it,jt'}\xi_j^{t'}\right]$$

We denote the second term by D_n. It now remains only to show that $E(D_n) = 0$. Note the following:

$$\begin{aligned}
E\left(1\left\{Z_i^t = P\right\}\xi_i^t\right) &= P(Z_i^t = P)E\left[Y_i^t(0) - \mu_0(X_i^{t-P},\ldots,X_i^{t-P-L}) \mid Z_i^t = P\right] \\
&= P(Z_i^t = P)E\left[Y_i^t(0) \mid Z_i^t = P\right] - P(Z_i^t = P)E\left[\mu_0(X_i^t,\ldots,X_i^{t-L}) \mid Z_i^t = P\right] \\
&= P(Z_i^t = P)E\left\{E\left[Y_i^t(0) \mid Z_i^t = P, (X_i^1,\ldots,X_i^T)\right] \mid Z_i^t = P\right\} \\
&\quad - P(Z_i^t = P)E\left[\mu_0(X_i^{t-P},\ldots,X_i^{t-P-L}) \mid Z_i^t = P\right] \\
&= P(Z_i^t = P)E\left\{E\left[Y_i^t(0) \mid Z_i^{t-P} = 0, (X_i^{t-P},\ldots,X_i^{t-P-L})\right] \mid Z_i^t = P\right\} \\
&\quad - P(Z_i^t = P)E\left[\mu_0(X_i^{t-P},\ldots,X_i^{t-P-L}) \mid Z_i^t = P\right] \\
&= P(Z_i^t = p)E\left[\mu_0(X_i^{t-P},\ldots,X_i^{t-P-L}) - \mu_0(X_i^{t-P},\ldots,X_i^{t-P-L}) \mid Z_i^t = P\right] = 0
\end{aligned}$$

The third equality follows from the law of iterated expectation, the fourth follows from the combination of L-ignorability and covariate L-exogeneity, and the last one from time-agnosticism. Finally, let $\mathbf{Z}$ be the vector of values Z_i^T for all i, and let $\mathbf{X}$ be the matrix of values X_i^t for all i, t. Then:

$$\begin{aligned}
&E\left(1\left\{Z_i^t = P\right\}1\left\{Z_j^{t'} = 0\right\}M_{it,jt'}\xi_j^{t'}\right) = E\left[1\left\{Z_i^t = P\right\}1\left\{Z_j^{t'} = 0\right\}E\left(M_{it,jt'}\xi_j^{t'} \mid \mathbf{Z}, Z_i^T = 0, \mathbf{X}\right)\right] \\
&= E\left[1\left\{Z_i^t = P\right\}1\left\{Z_j^{t'} = 0\right\}E\left(M_{it,jt'} \mid \mathbf{Z}, Z_i^T = 0, \mathbf{X}\right)E\left(\xi_j^{t'} \mid \mathbf{Z}, Z_i^T = 0, , \mathbf{X}\right)\right]
\end{aligned}$$

where the first line follows from iterated expectation and the second follows from the uniform exactness of the match. To finish the argument, note that

$$\begin{aligned}
E\left(\xi_j^{t'} \mid \mathbf{Z}, Z_i^T = 0, \mathbf{X}\right) &= E\left(Y_j^{t'}(0) \mid Z_i^T = 0, X_i^1,\ldots,X_i^T\right) - \mu_0(\mathbf{x}) \\
&= E\left(Y_j^{t'}(0) \mid Z_i^{t-P} = 0, (X_i^{t-P},\ldots,X_i^{t-P-L}) = \mathbf{x}\right) - \mu_0(\mathbf{x}) \\
&= \mu_0(\mathbf{x}) - \mu_0(\mathbf{x}) = 0
\end{aligned}$$

Here the first line follows from independent sampling of trajectories, the second from L-ignorability and covariate L-exogeneity, and the final one from time-agnosticism. Combining all the above facts and applying linearity of expectation suffices to prove $E(D_n) = 0$.

Proof of Lemma 1. $F(\mathbf{x})$ satisfies the capacity constraints of Network A since by feasibility of $\mathbf{x}$ for Problem A, we have $\sum_{i=1}^{N_1}\sum_{k=1}^{m_j} x_{ijk} \leq 1\ \forall j$. The flow conservation constraints at nodes κ_{jk} and κ_j are satisfied by construction. In addition the flow conservation constraints at τ_i are satisfied since by feasibility of $\mathbf{x}$ for Problem A, we have $\sum_{j=1}^{N_0}\sum_{k=1}^{m_j} x_{ijk} = w\ \forall i$. Similarly, the conservation constraints are satisfied at ω since by feasibility of $\mathbf{x}$ for Problem A, we have $\sum_{i=1}^{N_1}\sum_{j}^{N_0}\sum_{k=1}^{m_j} x_{ijk} = N_1 w$.

Proof of Lemma 2. By construction of G, $G(F(\mathbf{x})) = \mathbf{x}$ for any $\mathbf{x} \in \mathcal{F}_A^P$. Thus by Lemma 1, F is injective. To finish the proof, it suffices to show that G is also injective. Take any $\mathbf{x}', \mathbf{x}'' \in \mathcal{F}_A^N$ such that $G(\mathbf{x}') = G(\mathbf{x}'')$. By the definition of G, $x'_{(\tau_i,\kappa_{jk})} = x''_{(\tau_i,\kappa_{jk})}\ \forall i, j, k$. By conservation of flow at nodes κ_{jk}, we must also have $x'_{(\kappa_{jk},\kappa_j)} = \sum_{i=1}^{N_1} x'_{(\tau_i,\kappa_{jk})} = x''_{(\kappa_{jk},\kappa_j)}\ \forall j, k$. Finally, by conservation of at κ_j we have $x'_{(\kappa_j,\omega)} = \sum_{i=1}^{N_1}\sum_{k=1}^{m_j} x'_{(\kappa_j,\omega)} = x''_{(\kappa_j,\omega)}\ \forall j$. Therefore $\mathbf{x}' = \mathbf{x}''$ and G is an injective map also. Thus F is bijective with $G = F^{-1}$.

references

Abadie, A. and Imbens, G. W. (2006) Large sample properties of matching estimators for average treatment effects. *econometrica*, **74**, 235–267.

— (2012) A martingale representation for matching estimators. *Journal of the American Statistical Association*, **107**, 833–843.

Bertsekas, D. P. (1991) *Linear network optimization: algorithms and codes*. Cambridge, MA: MIT Press.

— (1998) *Network optimization: continuous and discrete models*. Belmont, MA: Athena Scientific.

Bertsekas, D. P., Tseng, P. et al. (1994) *RELAX-IV: A faster version of the RELAX code for solving minimum cost flow problems*. Massachusetts Institute of Technology, Laboratory for Information and Decision Systems Cambridge, MA.

Bohl, A. A., Fishman, P. A., Ciol, M. A., Williams, B., LoGerfo, J. and Phelan, E. A. (2010) 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*, **58**, 853–860.

Broaddus, M., Bailey, P. and Aron-Dine, A. (2018) Medicaid expansion dramatically increased coverage for people with opioid-use disorders, latest data show. *Retrieved from https://www.cbpp.org/sites/default/files/atoms/files/2-28-18health.pdf*.

Chen, A., Clarkwest, A., Croake, S., Felt-Lisk, S., Maxfield, M., Smith, L., Witmer, S., Zurovac, J., Lucado, J., McGivern, L., Paez, K. and Schur, C. (2011) Independent evaluation of the ninth scope of work qio program final report, volume i: Findings. *Tech. rep.*, Mathematica Policy Research.

Chew, R., Jones, K., Manley, M., Witman, A., Beadles, C., Liu, Y. and Larson, A. (2018) *rollmatch: Rolling Entry Matching*. URL: `https://CRAN.R-project.org/package=rollmatch`. R package version 1.0.1.

Dummit, L., Marrufo, G., Marshall, J., Ackerman, T., Bergman, S., Bradley, A. et al. (2016) Cms bundled payments for care improvement initiative models 2-4: Year 4 evaluation & monitoring annual report. *Tech. rep.*, The Lewin Group.

Hansen, B. (2007) Optmatch (r package optmatch). *R News*, **7**, 18–24.

Hansen, B. B. (2008) The prognostic analogue of the propensity score. *Biometrika*, 481–488.

Hansen, B. B. and Klopfer, S. O. (2006) Optimal full matching and related designs via network flows. *Journal of Computational and Graphical Statistics*, **15**, 609–627.

Haviland, A., Nagin, D. S. and Rosenbaum, P. R. (2007) Combining propensity score matching and group-based trajectory analysis in an observational study. *Psychological methods*, **12**, 247.

Hernán, M. A., Brumback, B. and Robins, J. M. (2001) Marginal structural models to estimate the joint causal effect of nonrandomized treatments. *Journal of the American Statistical Association*, **96**, 440–448.

Hill, I., Benatar, S., Courtot, B., Dubay, L., Blavin, F., Garrett, B., Howell, E., Allen, E., Cheeks, M., Thornburgh, S., Markell, J., Morgan, J. and Todd, H. (2018) Strong start for mothers and newborns evaluation: Year 4 annual report. *Tech. rep.*, Urban Institute.

Hill, J. and Reiter, J. P. (2006) Interval estimation for treatment effects using propensity score matching. *Statistics in Medicine*, **25**, 2230–2256.

Imai, K., Kim, I. S. and Wang, E. (2019) Matching methods for causal inference with time-series cross-section data. `http://web.mit.edu/insong/www/pdf/tscs.pdf`.

Imbens, G. W. and Wooldridge, J. M. (2009) Recent developments in the econometrics of program evaluation. *Journal of economic literature*, **47**, 5–86.

Kapelner, A. and Krieger, A. (2014) Matching on-the-fly: Sequential allocation with higher power and efficiency. *Biometrics*, **70**, 378–388.

Li, Y. P., Propert, K. J. and Rosenbaum, P. R. (2001) Balanced risk set matching. *Journal of the American Statistical Association*, **96**, 870–882.

Lu, B. (2005) Propensity score matching with time-dependent covariates. *Biometrics*, **61**, 721–728.

Ming, K. and Rosenbaum, P. R. (2001) A note on optimal matching with variable controls using the assignment algorithm. *Journal of Computational and Graphical Statistics*, **10**, 455–463.

Papadimitriou, C. H. and Steiglitz, K. (1982) *Combinatorial optimization: algorithms and complexity*. Courier Corporation.

Peikes, D., Anglin, G., Dale, S., Taylor, E. F., O'Malley, A., Ghosh, A., Swankoski, K., Crosson, J., Keith, R., Mutti, A., Hoag, S., Singh, P., Tu, H., Grannemann, T., Finucane, M., Zutji, A., Vollmer, L., Brown, R. et al. (2018) Evaluation of the comprehensive primary care initiative: Third annual report. *Tech. rep.*, Mathematica Policy Research.

Pimentel, S. D., Kelz, R. R., Silber, J. H. and Rosenbaum, P. R. (2015a) 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*, **110**, 515–527.

Pimentel, S. D., Yoon, F. and Keele, L. (2015b) Variable-ratio matching with fine balance in a study of the peer health exchange. *Statistics in medicine*, **34**, 4070–4082.

Robins, J. M. (2000) Marginal structural models versus structural nested models as tools for causal inference. In *Statistical models in epidemiology, the environment, and clinical trials*, 95–133. Springer.

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

— (1989) Optimal matching for observational studies. *Journal of the American Statistical Association*, **84**, 1024–1032.

— (2002) *Observational Studies*. New York, NY: Springer.

— (2010) *Design of Observational Studies*. New York, NY: Springer.

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

— (1985) Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician*, **39**, 33–38.

Rubin, D. B. (2007) The design versus the analysis of observational studies for causal effects: parallels with the design of randomized trials. *Statistics in medicine*, **26**, 20–36.

Stuart, E. A. (2010) Matching methods for causal inference: A review and a look forward. *Statistical science: a review journal of the Institute of Mathematical Statistics*, **25**, 1.

Witman, A., Beadles, C., Liu, Y., Larsen, A., Kafali, N., Gandhi, S., Amico, P. and Hoerger, T. (2019) Comparison group selection in the presence of rolling entry for health services research: Rolling entry matching. *Health Services Research*, **54**, 492–501.

Yang, D., Small, D. S., Silber, J. H. and Rosenbaum, P. R. (2012) Optimal matching with minimal deviation from fine balance in a study of obesity and surgical outcomes. *Biometrics*, **68**, 628–636.

Zubizarreta, J. R., Small, D. S. and Rosenbaum, P. R. (2014) Isolation in the construction of natural experiments. *The Annals of Applied Statistics*, 2096–2121.

— (2018) A simple example of isolation in building a natural experiment. *Chance*, **31**, 16–23.