Efficient algorithms for building representative matched pairs with enhanced generalizability

Bo Zhang

Vaccine and Infectious Disease Division, Fred Hutchinson Cancer Center, Seattle, Washington, USA

Abstract

Many recent efforts center on assessing the ability of real-world evidence (RWE) generated from non-randomized, observational data to produce results compatible with those from randomized controlled trials (RCTs). One noticeable endeavor is the RCT DUPLICATE initiative. To better reconcile findings from an observational study and an RCT, or two observational studies based on different databases, it is desirable to eliminate differences between study populations. We outline an efficient, network-flow-based statistical matching algorithm that designs well-matched pairs from observational data that resemble the covariate distributions of a target population, for instance, the target-RCT-eligible population in the RCT DUPLICATE initiative studies or a generic population of scientific interest. We demonstrate the usefulness of the method by revisiting the inconsistency regarding a cardioprotective effect of the hormone replacement therapy (HRT) in the Women’s Health Initiative (WHI) clinical trial and corresponding observational study. We found that the discrepancy between the trial and observational study persisted in a design that adjusted for the difference in study populations’ cardiovascular risk profile, but seemed to disappear in a study design that further adjusted for the difference in HRT initiation age and previous estrogen-plus-progestin use. The proposed method is integrated into the R package match2C.

KEYWORDS
generalizability, RCT DUPLICATE initiative, statistical matching, trial emulation, Women’s health initiative

1 | INTRODUCTION

1.1 Comparing observational studies to randomized controlled trials

In a recent high-profile study published in Circulation, the RCT DUPLICATE initiative (Franklin et al., 2020, 2021) designed 10 observational studies using retrospective, non-randomized claims data and compared their real-world-evidence-based (RWE-based) treatment effect estimates to those based on randomized controlled trials (RCTs) investigating very similar clinical questions. To facilitate a more informed comparison of effect estimates obtained from observational and RCT data, it is desirable to design an observational study whose treated and matched control groups are comparable to the RCT population in baseline covariates. In their design stage, Franklin et al. (2021) carefully emulated the RCT population by applying the same inclusion and exclusion criteria to the observational data prior to statistical matching; however, tangible
and potentially meaningful differences persist. Franklin et al. (2021) concluded:

[Int] Inclusion and exclusion criteria from the trials could only be partially emulated, and even where fully emulated, the resulting distributions were at times meaningfully different between the RCT and RWE populations, possibly because of nonrepresentative participation in RCTs.

Discrepancies in age, race/ethnicity and important pre-existing comorbid conditions between RCT and observational study populations are common in Franklin et al.'s (2021) emulation studies. For instance, the Saxagliptin and Cardiovascular Outcomes in Patients with Type 2 Diabetes Mellitus (SAVOR-TIMI 53) study, one of the ten emulated trials, enrolled 33.1% female participants and 37.8% participants with history of myocardial infarction (MI); the observational study emulating the trial, however, consisted of 46.8% female participants and 11.2% participants with MI (Franklin et al., 2021, Table 1). These differences persisted in the final matched cohort and could at least partially explain the disagreement in effect estimates.

The ongoing RCT emulation study led by the RCT DUPLICATE initiative and many similar endeavors to better reconcile RCT and observational study findings, for instance, the Women’s Health Initiative (WHI) study (Hernán et al., 2008; Prentice et al., 2005), or findings derived from different observational databases, motivate us to develop a transparent, efficient, and easy-to-use algorithm that constructs homogeneous matched pairs that resemble a template (i.e., a random sample from the target population) in key covariates.

There are two important applications for such an algorithm. First, in the context of target trial emulation using observational data (Franklin et al., 2021), the target population would be the RCT-eligible population, the template would consist of trial participants, and the algorithm could be used to construct treated and matched control groups resembling the target RCT population. Second, the algorithm may be used to help assess inconsistencies often seen among observational studies based on different databases. For instance, in a recent study of the impact of intraoperative transesophageal echocardiography (TEE) on patients’ clinical outcomes, Metkus et al. (2021) conducted a matched cohort study and found an effect smaller than that derived by MacKay et al. (2021). Metkus et al.’s (2021) analysis was based on the Society of Thoracic Surgeons Adult Cardiac Surgery Database (STS ACSD), while MacKay et al.’s (2021) analysis was based on Medicare beneficiaries at higher risk compared to the adult population in the STS ACSD. In this example, Medicare beneficiaries would be the target population, the template would consist of a random sample from Medicare claims data, and our algorithm could be used to construct matched cohorts from the STS ACSD that mimic Medicare beneficiaries in risk factors. This could enable researchers to estimate the effect of TEE for an important population (elderly Americans enrolled in Medicare) using STS ACSD, the world’s premier clinical outcomes registry for adult cardiac surgery, and facilitate a more informed comparison between results derived from different databases.

1.2 | A two-step method: a new approach—an important caveat

One strategy to create homogeneous matched pairs resembling a template would consist of two steps. Take the RCT DUPLICATE initiative as an example. In the first step, researchers could select a subset of treated participants from the observational data via matching on covariates collected by both the RCT and the observational study. In the second step, controls in the observational database are then matched to the selected treated group. Each step involves only two groups and could be done using standard statistical matching algorithms like the network-flow-based algorithms built upon bipartite graphs (Rosenbaum, 1989, 2010; Silber et al., 2014). A graph is bipartite if its vertices could be divided into two disjoint sets such that no two vertices in the same set are adjacent; see Section 2.1 for a review. Such a two-step approach was adopted in the context of building evidence factors and forming matched triplets (Karmakar et al., 2019). Alternatively, researchers could adopt a mixed-integer-programming-based (MIP-based) approach (Bennett et al., 2020; Zubizarreta, 2012).

This two-step approach suffers from a major drawback. Franklin et al. (2021) matched on more than 120 covariates collected by the claims data to guard against unmeasured confounding in their emulation of the target trial, although the trial reported only around 20 baseline covariates. In the first step, there are many ways to design a smaller treated group similar to the RCT-eligible population in 20 covariates; however, it is difficult to determine which treated group designed from the first step should be used to form the final match. A selected treated group similar to the RCT-eligible population in 20 covariates could have poor overlap with controls in the observational database in the other more than 100 covariates; see Web Appendix C.1 for an illustration. Moreover, resemblance of the matched cohort to a target population is useful but not necessary in many applications; priority should be given to creating well-balanced matched cohorts to first maximize a study’s internal validity. The two-step approach therefore
precludes a principled trade-off between a matched cohort study’s internal validity and its generalizability. In this paper, “generalizability” always refers to generalizability to a target population and should not be confused with notions of “real-world efficacy” in populations to which the treatment might be deployed.

In principle, Bennett et al.’s (2020) MIP-based method could be adapted to designing matched treated and control groups well balanced in many covariates while resembling a template by modifying constraints in the mathematical program. Although recent advancements in computing power have made it more practical to solve large-scale MIP problems with complex constraints, the general MIP problems are still theoretically intractable or NP-hard. From a practical perspective, MIP-based methods require installing a powerful commercial optimization routine (e.g., Gurobi), which is proprietary and could be an obstacle to researchers. Other approaches to adjusting for sample selection bias include weighting and doubly-robust methods (see, e.g., Stuart et al., 2011; Dahabreh et al., 2019).

Compared to MIP-based methods, network-flow-based methods only require solving a polynomial solvable problem that is tractable both in theory and practice and have proven successful in empirical comparative effectiveness research for decades (Rosenbaum, 2002, 2010; Stuart, 2010). The primary goal of this paper is to outline an efficient, network-flow-based algorithm that constructs matched pairs from observational data with close resemblance to a template from a target population. Our primary insight is that, compared to solving optimization problems based on a bipartite graph multiple times, a better strategy is to solve an optimization problem built upon a properly designed, tripartite network only once. The proposed network structure is referred to as tripartite because it contains three parts, instead of two, as in a bipartite network. We demonstrate the usefulness of the proposed method by revisiting the WHI study and exploring how our method facilitates different study designs and yields insight into the inconsistency between the WHI observational study and trial findings. We found that the discrepancy regarding a cardioprotective effect of the hormone replacement therapy (HRT) persisted in a design that adjusted for the difference in the cardiovascular risk profile differences between the observational and trial data, but seemed to disappear in a design that further adjusted for the difference in HRT initiation age and previous estrogen-plus-progestin use, lending some evidence-based support to Willett et al.’s (2006) arguments.

We conclude this section with an important caveat. Many reasons may contribute to the disagreement between effect estimates derived from an observational study and an RCT or from two observational studies. The interaction between effect heterogeneity and difference in covariate distributions is only one of the many possible reasons. Other important reasons include unmeasured confounding, noncompliance in the RCT, difference in the definition of treatment and/or clinical endpoints, among others. A matched cohort constructed from observational data, no matter how balanced its treated and matched control groups are, and how similar these two groups are to the target population, is not necessarily free from unmeasured confounding bias and, in our opinion, cannot replace an RCT. Nevertheless, addressing the difference in covariate distributions across studies helps researchers focus on other explanations of the inconsistency in the findings. A better understanding of the inconsistency would ultimately help researchers better understand the underlying mechanisms.

2 | METHODOLOGY

2.1 | Network-flow-based matching algorithms

Matching and subclassification are widely used in empirical research to embed non-randomized, observational data into an approximate RCT (Ho et al., 2007; Rosenbaum, 2002, 2010; Stuart, 2010). In a seminal paper, Rosenbaum (1989) first bridged statistics literature on constructing matched samples and operations research and computer science literature on matching in graphs and networks. Rosenbaum’s (1989) key insight is to represent treated and control units as vertices in a bipartite graph and recast the statistical problem of constructing a matched cohort as an optimization problem of finding a minimum-cost network flow, a standard combinatorial optimization problem for which efficient algorithms exist. Figure 1 illustrates this paradigm using a small example of three treated and

![Network-flow representation of a small matching problem. Five control units C₁, C₂, ..., C₅ are to be matched to three treated units T₁, T₂, and T₃. Each edge connecting Tᵢ, i = 1, 2, 3, and Cⱼ, j = 1, 2, 3, 4, 5, is associated with a flow capacity (equal to 1 in pair matching) and a cost equal to a pre-specified covariate distance between Tᵢ and Cⱼ.](image-url)
five control units. Readers should imagine that the source node on the far left emits three units of flow, one arriving at the node $T_1$, one at $T_2$ and the other at $T_3$. Edges of the form $\{e(T_i, C_j)\}$ connecting $T_i$ and $C_j$, $i = 1, 2, 3$ and $j = 1, ..., 5$, each have capacity 1 and are associated with a cost $\text{cap}(e(T_i, C_j))$ equal to some pre-specified covariate distance between $T_i$ and $C_j$. These three units of flow at nodes $T_1$, $T_2$, and $T_3$ then choose to flow in a way that minimizes the total cost. To illustrate, Figure 1 displays a minimum-cost flow (bold, black lines) consisting of edges $\{e(T_1, C_1), e(T_2, C_2), e(T_3, C_4)\}$. In this way, three matched pairs $\{(T_1, C_3), (T_2, C_1), (T_3, C_4)\}$ are constructed. Network-flow-based matching algorithms do not allow reusing the control units. Recent developments have also modified the network structure displayed in Figure 1 to handle large datasets or to incorporate additional design features; see Web Appendix A for a brief review. Network-flow-based algorithms based on bipartite networks have also been utilized to build matched triplets (Karmakar et al., 2019; Nattino et al., 2021). Recently, a tripartite structure was proposed in the literature to match one sample according to two criteria (Zhang et al., 2021). Our proposed method utilizes the idea of a tripartite network to build representative matched pairs.

2.2 Basic network structure: vertices and edges

We describe our proposed method using the RCT emulation study described in Section 1.1 for illustration purpose; the method accommodates an arbitrary template from a target population other than the RCT-eligible population (e.g., the population of Medicare beneficiaries as discussed in the TEE and cardiac surgery example in Section 1.1).

Figure 2 depicts a basic version of the proposed network structure. We will refer to study participants from the RCT as RCT units and those from the observational database as OBS units for short. There are $R$ units from the target RCT. These $R$ RCT units $\mathcal{K} = \{k_1, ..., k_R\}$ are represented by nodes labeled $k_r$, $r = 1, ..., R$. There are $T \geq R$ OBS treated units and $C \geq T$ OBS control units from an administrative database. OBS treated units $\mathcal{T} = \{t_1, ..., t_T\}$ are represented twice in the network, by nodes labeled $t_t$ and $\bar{t}_t$, $t = 1, ..., T$, and OBS control units $\mathcal{C} = \{c_1, ..., c_C\}$ are represented by nodes labeled $c_c$, $c = 1, ..., C$. In addition to $R + 2T + C$ nodes representing RCT and OBS study participants, there is a source node $\xi$ and a sink node $\bar{\xi}$, so that the network consists of $|\mathcal{V}| = R + 2T + C + 2 = O(C)$ nodes in total:

$$\mathcal{V} = \left\{\xi, k_1, ..., k_R, t_1, ..., t_T, \bar{t}_1, ..., \bar{t}_T, c_1, ..., c_C, \bar{\xi}\right\}.$$

An ordered pair of vertices is referred to as an edge in the network. The basic structure in Figure 2 consists of the following edges:

$$\mathcal{E} = \left\{\{(\xi, k_r), (k_r, t_r), (t_r, \bar{t}_r), (\bar{t}_r, c_c), (c_c, \bar{\xi})\} : r = 1, ..., R, t = 1, ..., T, c = 1, ..., C\right\}.$$

There are a total of $|\mathcal{E}| = O(C^2)$ edges assuming $C$ is a constant multiple of $T$.

2.3 Basic network structure: capacity and matched pairs

To transform a statistical matching problem into an appropriate network-flow-based optimization problem, one needs to carefully design the cost and capacity of each edge. Fix an integer $k$ such that $1 \leq k \leq \lceil T/R \rceil$ and consider constructing $N = kR$ matched pairs from the observational database. For instance, the SAVOR-TIMI 53 trial consists of $R = 8,280$ participants assigned saxagliptin, the intervention under evaluation, and three retrospective databases (Optum, MarketScan, and Medicare) consist of $T = 91,082$ participants using saxagliptin according to Franklin et al.’s (2021) study protocol registered at ClinicalTrials.gov (identifier NCT03936023). Researchers could in principle choose any integer $k$ between 1 and $91,082/8,280 = 11$ in this example.

Let $\text{cap}(e) \geq 0$ denote the capacity of an edge $e \in \mathcal{E}$. In the basic network structure depicted in Figure 2, $\text{cap}(\{(\xi, k_r)\}) = k$ for $r = 1, ..., R$, and all other edges have capacity 1. In Figure 2, the source $\xi$ supplies $k R$ units of flow, the sink $\bar{\xi}$ absorbs $k R$ units of flow, while all other nodes preserve the flow by simply passing them along (Rosenbaum, 1989). A feasible flow $f(\cdot)$ of the proposed network is formally defined as a mapping from the set of edges $\mathcal{E}$ to $\{0, ..., k\}$ such that (i) all capacity constraints are respected, that is, $0 \leq f(e) \leq \text{cap}(e)$, $e \in \mathcal{E}$, (ii) $kR$ units of flow are supplied at $\xi$ and absorbed at $\bar{\xi}$, that is, $\sum_{r=1}^{R} f(\{(\xi, k_r)\}) = kR$ and $\sum_{c=1}^{C} f(\{(c, \bar{\xi})\}) = kR$, and (iii) the flow is preserved at all nodes but $\xi$ and $\bar{\xi}$, that is, $\sum_{(a,b)\in\mathcal{E}\setminus\{(\xi, k_r)\}} f(\{(a, b)\}) = \sum_{(b,c)\in\mathcal{E}\setminus\{(c, \bar{\xi})\}} f(\{(b, c)\})$ for all $b \in \mathcal{V} \setminus \{\xi, \bar{\xi}\}$.

It is beneficial to consider a concrete example. The toy example in Figure 2 has $R = 3$, $T = 4$, and $C = 6$. Consider setting $k = 1$ so that all edges in the network have capacity 1. Thick, black lines in Figure 2 correspond to one (out of $\binom{3}{2} \times 6 \times 5 \times 4 = 480$) feasible flows in this network. The left part of the network helps select OBS treated units using RCT units as a template, while the right part of the network performs...
the actual statistical matching and returns a matched sample. Formally, the matched sample $M$ is defined by $M = \{(\tau_c, \gamma_c)\}$ such that $f(\tau_c, \gamma_c) = f(\tau_t, \gamma_t) = 1$. For instance, the matched sample returned by the feasible flow in Figure 2 consists of $M = \{(\tau_1, \gamma_1), (\tau_2, \gamma_3), (\tau_3, \gamma_4)\}$.

### 2.4 Basic network structure: cost, probability of participation, propensity score

While the network infrastructure, including vertices, edges, and capacities, determines the collection of feasible flows, costs associated with each edge help select one best suited for empirical researchers’ specific purposes. Let $c(e)$ denote the cost associated with edge $e \in \mathcal{E}$. In the basic network structure depicted in Figure 2, we let $c(\xi, \tau_t) = c(\tau_t, \tau_t) = c(\gamma_c, \xi) = 0$ for $r = 1, \ldots, R$, $t = 1, \ldots, T$, and $c = 1, \ldots, C$.

Costs associated with edges $(\xi, \tau_t)$ play a key role in forcing selected OBS units to resemble the template. Suppose that each RCT unit $\xi$ is associated with a vector of covariates $\mathbf{x}$ and each OBS unit is associated with $(\mathbf{\xi}, \mathbf{x})$. As in the RCT DUPLICATE example, $\mathbf{x}$ contains roughly 20 covariates that both the RCT and observational database collected, while $\mathbf{\xi}$ consists of more than 100 additional covariates available only in the claims database. Let $\delta_{\xi,\tau_t}(\mathbf{\xi}, \mathbf{x})$ denote a measure of covariate distance between $\xi$ and $\tau_t$ in $\mathbf{\xi}$. The $\delta_{\xi,\tau_t}(\mathbf{\xi}, \mathbf{x})$ is equal to $\delta_{\xi,\tau_t}(\mathbf{\xi}, \mathbf{x}) = 0$. One intriguing strategy equals $\delta_{\xi,\tau_t}(\mathbf{\xi}, \mathbf{x})$ to a scalar “balancing score” of $\mathbf{x}$ (Rosenbaum and Rubin, 1983). In the context of generalizing an RCT’s effect estimates to a target population, Stuart et al. (2011) studied the conditional probability of selecting into the RCT, which is a balancing score. Matching on this probability helps stochastically balance covariates $\mathbf{x}$ used to construct this probability. Variables available in both the template and the observational database, and are believed to potentially confound the treatment assignment and the outcome of interest should be included in $\mathbf{x}$ when estimating this probability via a logistic regression (Stuart et al., 2011) or some other flexible estimation procedure.

Lastly, let $\Delta_{\tau_i,\gamma_c}(\mathbf{\xi}, \mathbf{x})$ denote a measure of distance between OBS treated unit $\tau_i$ and OBS control unit $\gamma_c$ in their observed covariates $(\mathbf{\xi}, \mathbf{x})$. The $\Delta_{\tau_i,\gamma_c}(\mathbf{\xi}, \mathbf{x})$ is equal to $\Delta_{\tau_i,\gamma_c}(\mathbf{\xi}, \mathbf{x}) = 0$. One widely used strategy equals $\Delta_{\tau_i,\gamma_c}(\mathbf{\xi}, \mathbf{x})$ to the absolute difference in the propensity score (Rosenbaum & Rubin, 1983). Alternatively, one may let $\Delta_{\tau_i,\gamma_c}(\mathbf{\xi}, \mathbf{x})$ be the Mahalanobis distance in $(\mathbf{\xi}, \mathbf{x})$ or the Mahalanobis distance within a propensity score caliper (Rosenbaum & Rubin, 1985). Rosenbaum and Rubin (1985) recommended matching on the Mahalanobis distance within calipers defined by the estimated propensity scores.
2.5 Minimum cost flow, complexity, trade-off between validity and generalizability

The cost of a feasible flow \( f(\cdot) \) in the proposed network is equal to

\[
\text{cost}(f) = \sum_{e \in E} f(e) \cdot \text{cost}(e) = \sum_{(k_r, \tau_t) \in E: f((k_r, \tau_t)) = 1} \delta_{k_r, \tau_t}(\mathbf{x}) \tag{S1}
\]

\[
+ \sum_{(\tau_t, \gamma_c) \in E: f((\tau_t, \gamma_c)) = 1} \Delta_{\tau_t, \gamma_c}(\mathbf{x}, \mathbf{x}) \tag{S2}
\]

and a flow \( f(\cdot) \) is a minimum cost flow if it is feasible, and every other feasible flow has a cost at least as high as \( \text{cost}(f) \). A minimum cost flow can be found in \( O(|V| \cdot |E| + |V|^2 \log(|V|)) \) operations (Korte & Vygen, 2011) and in the proposed network, we have \(|V| = O(C)\), \(|E| = O(C^2)\), so the computation complexity \( O(|V| \cdot |E| + |V|^2 \log(|V|)) \) simplifies to \( O(C^3) \).

There is a tension between a matched cohort study’s internal validity and its generalizability to a target population. According to our formulation, cost \( S_2 = \sum_{(\tau_t, \gamma_c) \in E: f((\tau_t, \gamma_c)) = 1} \Delta_{\tau_t, \gamma_c}(\mathbf{x}, \mathbf{x}) \) in expression (1) measures the homogeneity between the matched treated and control groups. A small \( S_2 \) value reflects the observational data being well-matched and enhances the validity of the matched cohort study. On the other hand, the cost \( S_1 = \sum_{(k_r, \tau_t) \in E: f((k_r, \tau_t)) = 1} \delta_{k_r, \tau_t}(\mathbf{x}) \) measures how well matched pairs resemble the target population, and a small \( S_1 \) indicates matched cohort study’s improved generalizability. To explore the trade-off between a matched cohort study’s internal validity and its generalizability to a target population, we replace \( \Delta_{\tau_t, \gamma_c}(\mathbf{x}, \mathbf{x}) \) with \( \lambda \cdot \Delta_{\tau_t, \gamma_c}(\mathbf{x}, \mathbf{x}) \) for some \( \lambda > 0 \) so that \( \text{cost}(f) \) becomes:

\[
\text{cost}(f) = \sum_{e \in E} f(e) \cdot \text{cost}(e) = \sum_{(k_r, \tau_t) \in E: f((k_r, \tau_t)) = 1} \delta_{k_r, \tau_t}(\mathbf{x})
\]

\[
+ \lambda \cdot \sum_{(\tau_t, \gamma_c) \in E: f((\tau_t, \gamma_c)) = 1} \Delta_{\tau_t, \gamma_c}(\mathbf{x}, \mathbf{x}) \tag{2}
\]

According to this formulation, a large \( \lambda \) value prioritizes a matched cohort study’s internal validity, while a small \( \lambda \) value prioritizes its generalizability to the target population. A similar weighting scheme is also used in Zhang et al. (2021) but for a different purpose.

2.6 Practical considerations and software availability

The proposed matching algorithm is integrated in the package match2C as function template_match available via the statistical computing software R with a detailed tutorial.

To successfully implement the method, users need to specify two parameters. Parameter \( k \) controls the size of the matched cohort and is determined by the size of the template, size of the observational database, overlap of the template, and observational database in covariate distributions and power considerations. In general, a small \( k \) corresponds to constructing a small matched cohort and this small matched cohort by design will be more closely matched and better resemble the template compared to a larger matched cohort corresponding to a larger \( k \) value; see the illustrative example in Web Appendix C.2 and simulation studies in Section 3. If researchers find the matched comparison not adequately powered, then larger \( k \) values should be explored. Parameter \( \lambda \) controls the trade-off between the validity of the matched comparison and its generalizability to the target population. A large \( \lambda \) prioritizes the internal validity of the matched cohort study. When the treated and control groups in the observational database are well-overlapped so that there are many possible internally-valid matched cohort studies, researchers may reduce \( \lambda \) to further improve the matched cohort study’s generalizability; see Web Appendix F for a detailed tutorial. As the parameter \( \lambda \) controls the relevant contribution of costs \( S_1 \) and \( S_2 \) to the overall cost, its magnitude depends on the relative magnitude of \( S_1 \) and \( S_2 \). Depending on the number of edges of the form \((k_r, \tau_t)\) and \((\tau_t, \gamma_c)\) and the distance metrics, \( S_1 \) and \( S_2 \) could differ in multiple orders of magnitude; hence, we recommend varying \( \lambda \) by at least a factor of 10 each time.

Matching is part of the design of an observational study and should be carried out without looking at the outcome data. Good practice includes keeping time-stamped analysis logs for review and posting a detailed pre-analysis protocol; see, for example, Franklin et al. (2020, 2021). Provided that no outcome data are viewed, researchers typically perform matching multiple times and select the design based on covariate balance. Recently, many formal
diagnostic tests of covariate balance have been proposed (Chen & Small, 2022; Gagnon-Bartsch & Shem-Tov, 2019; Yu, 2021). Researchers could perform a formal diagnostic test to assess if there is any residual imbalance in observed covariates ($\mathbf{x}$, $\mathbf{x}'$) between two groups in the matched cohort; analogously, a formal test could be carried out to examine residual imbalance in common covariates $\mathbf{x}$ between the matched group and the template. We will discuss these aspects more concretely when examining the WHI study in Section 4.

3 | SIMULATION STUDY

3.1 | Goal, structure, and measurement of success

Our primary goal in this section is to examine how the study designs delivered by different matching algorithms affect the bias. We are interested in the case where the effect is heterogeneous and effect modifiers’ distributions vary in the observational database and the target population. We considered the following target population: $\mathbf{X} \sim \text{multivariate normal}(\mu, \Sigma)$ with $\mu = (0.25, 0, 0, 0, 0)^T$ and $\Sigma = \mathbf{I}_{3 \times 5}$, and generated a template $\mathcal{K}$ consisting of a random sample of size 300 from the target population. We consider an observational database with $|T| = 1,000$ treated and $|C| = 3,000$ control units. The data-generating process for units in the observational database and statistical matching procedures to be investigated are specified via the following factorial design:

Factor 1: Dimension of covariates in the observational database, $d$: 10, 30, and 50.

Factor 2: Overlap, $\theta$: $\mathbf{X} \sim \text{Multivariate Normal}(\mu, \Sigma)$, with $\mu = (\beta_0, 0, \ldots, 0)^T$ and $\Sigma = \mathbf{I}_{d \times d}$. We consider $\theta = 0.50$ and 0.75.

Factor 3: Matching algorithms to be investigated, $\mathcal{M}$:

1. $\mathcal{M}_{\text{opt}}$: matching according to two criteria (Zhang et al., 2021): (i) minimizing the earthmover’s distance, a measure of distance between two probability distributions, between the distributions of the estimated propensity scores in the treated and matched control groups, and (ii) minimizing the within-matched-pair robust Mahalanobis distances. Algorithm $\mathcal{M}_{\text{opt}}$ produces 1,000 matched pairs and does not make use of the template.

2. $\mathcal{M}_{\text{template}, k=1}$: matching according to the proposed network structure (Figure 2) with $k = 1$. Algorithm $\mathcal{M}_{\text{template}, k=1}$ produces 300 matched pairs.

3. $\mathcal{M}_{\text{template}, k=2}$: similar to $\mathcal{M}_{\text{template}, k=1}$ but with $k = 2$. Algorithm $\mathcal{M}_{\text{template}, k=2}$ produces 600 matched pairs.

Factor 4: Tuning parameter in $\mathcal{M}_{\text{template}, k, \lambda}$: $100, 1, 0.01$.

Factors 1 and 2 define the data-generating process for units in the observational database. This data-generating process is motivated by applications discussed in Section 1.2, where the observational data (treated and control) are from a database more enriched than the database from which the template is sampled. Factors 3 and 4 define a total of $1 + 2 \times 3 = 7$ matching algorithms to be investigated. As discussed in Section 2.5, the tuning parameter $\lambda$ controls the trade-off between a matched cohort study’s validity and its generalizability to the target population.

For each unit, we further generate two potential outcomes $Y(0) \sim \text{Normal}(\mu|\lambda, \Sigma)$ and $Y(1) = Y(0) + \beta(X_1)$, and the observed outcome satisfies $Y = Z \cdot Y(1) + (1 - Z) \cdot Y(0)$. Factor 5 specifies the mean of the potential outcome $Y(0)$ and Factor 6 specifies the treatment effect:

Factor 5: Mean of $Y(0), X_1^T \nu$: a constant $\nu$ vector with all entries equal to 0, 0.05 or 0.1.

Factor 6: Treatment effect, $\beta(X_1)$: a constant effect $\beta(X_1) = 2$, a mildly heterogeneous effect $\beta(X_1) = 2 - 0.2X_1$, and a strongly heterogeneous effect $\beta(X_1) = 2 - X_1$.

The average treatment effect of the target population satisfies $\text{ATE}_{\text{target}} = 2$ when $\beta(X_1) = 2$, $\text{ATE}_{\text{target}} = 1.95$ when $\beta(X_1) = 2 - 0.2X_1$, and $\text{ATE}_{\text{target}} = 1.75$ when $\beta(X_1) = 2 - X_1$.

There are multiple ways to analyze matched-pair data. Examples include a parametric $t$-test, randomization inference (Rosenbaum, 2002, 2010) and regression adjustment (Ho et al., 2007; Rubin, 1979). In this simulation study, we calculated a difference-in-means estimator for matched data produced by each of the seven algorithms and compared these seven effect estimates to the target parameter $\text{ATE}_{\text{target}}$.

3.2 | Simulation results

Table 1 summarizes the percentage of bias of each difference-in-means estimator when the overlap parameter $\theta = 0.50$. Web Appendix D.1 reports similar results for $\theta = 0.75$.

We have observed a few consistent trends. First, for a constant treatment effect, the bias is relatively small under all data-generating processes and statistical matching algorithms under consideration. Second, when the treatment effect is heterogeneous and the effect modifier $X_1$ has a
| Heterogeneity Level $\mathcal{H}(\mathcal{X}_1)$ | DGP of $Y(0) \nu$ | $\mathcal{M}_{opt}$ | $k = 1 \lambda = 0.01$ | $k = 1 \lambda = 1$ | $k = 1 \lambda = 100$ | $k = 2 \lambda = 0.01$ | $k = 2 \lambda = 1$ | $k = 2 \lambda = 100$ |
|---------------------------------------------|-----------------|-------------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|
| $d = 10$                                    |                 |                   |                 |                 |                 |                 |                 |                 |
| Constant                                    | 0               | 0.07%             | 0.21%           | -0.06%          | -0.06%          | -0.01%          | 0.05%           | -0.06%          |
|                                            | 0.05            | 0.00%             | 0.33%           | 0.18%           | 0.14%           | 0.23%           | 0.25%           | 0.44%           |
|                                            | 0.10            | 0.06%             | 0.59%           | 0.30%           | 0.44%           | 0.73%           | 0.66%           | 0.61%           |
| Mild                                        | 0               | -5.17%            | -1.23%          | -1.19%          | -2.04%          | -1.52%          | -1.56%          | -2.56%          |
|                                            | 0.05            | -4.80%            | -0.11%          | -0.15%          | -1.25%          | -0.53%          | -0.65%          | -1.77%          |
|                                            | 0.10            | -5.00%            | 0.04%           | -0.20%          | -1.22%          | -0.58%          | -0.48%          | -1.44%          |
| Strong                                      | 0               | -24.99%           | -4.01%          | -3.75%          | -9.05%          | -6.45%          | -5.96%          | -11.68%         |
|                                            | 0.05            | -25.18%           | -3.10%          | -2.92%          | -8.40%          | -5.50%          | -5.60%          | -11.43%         |
|                                            | 0.10            | -25.23%           | -3.68%          | -3.39%          | -8.07%          | -6.14%          | -5.91%          | -11.25%         |
| $d = 30$                                    |                 |                   |                 |                 |                 |                 |                 |                 |
| Constant                                    | 0               | 0.23%             | 0.09%           | 0.45%           | 0.40%           | 0.31%           | 0.33%           | 0.28%           |
|                                            | 0.05            | 0.13%             | 1.04%           | 1.27%           | 1.37%           | 1.13%           | 1.19%           | 1.29%           |
|                                            | 0.10            | 0.12%             | 1.55%           | 1.43%           | 1.62%           | 1.70%           | 1.61%           | 1.83%           |
| Mild                                        | 0               | -4.77%            | -0.57%          | -1.32%          | -2.06%          | -1.07%          | -1.76%          | -2.50%          |
|                                            | 0.05            | -5.07%            | -0.55%          | -0.83%          | -1.56%          | -0.67%          | -1.27%          | -1.88%          |
|                                            | 0.10            | -4.63%            | 0.99%           | 0.80%           | 0.48%           | 0.62%           | 0.06%           | -0.41%          |
| Strong                                      | 0               | -24.71%           | -3.57%          | -5.90%          | -9.28%          | -5.65%          | -9.50%          | -12.83%         |
|                                            | 0.05            | -24.89%           | -3.13%          | -5.44%          | -8.80%          | -5.30%          | -8.96%          | -12.37%         |
|                                            | 0.10            | -24.78%           | -2.49%          | -4.88%          | -8.01%          | -4.51%          | -8.21%          | -11.35%         |
| $d = 50$                                    |                 |                   |                 |                 |                 |                 |                 |                 |
| Constant                                    | 0               | 0.06%             | 0.31%           | -0.21%          | -0.21%          | 0.19%           | -0.12%          | -0.06%          |
|                                            | 0.05            | 0.07%             | 0.77%           | 1.15%           | 1.12%           | 0.99%           | 1.29%           | 1.29%           |
|                                            | 0.10            | -0.05%            | 1.33%           | 1.82%           | 2.07%           | 1.66%           | 2.02%           | 2.18%           |
| Mild                                        | 0               | -5.12%            | -1.02%          | -1.59%          | -2.20%          | -1.60%          | -2.43%          | -2.94%          |
|                                            | 0.05            | -5.15%            | 0.37%           | -0.69%          | -1.26%          | -0.37%          | -1.16%          | -1.63%          |
|                                            | 0.10            | -5.03%            | 0.78%           | 0.20%           | -0.24%          | 0.47%           | 0.01%           | -0.40%          |
| Strong                                      | 0               | -24.64%           | -3.48%          | -6.82%          | -9.48%          | -5.84%          | -10.26%         | -12.72%         |
|                                            | 0.05            | -25.00%           | -3.22%          | -6.76%          | -9.06%          | -5.39%          | -9.88%          | -12.21%         |
|                                            | 0.10            | -24.78%           | -1.80%          | -5.01%          | -7.37%          | -3.97%          | -8.34%          | -10.63%         |

*Note:* The overlap parameter $\theta$ in Factor 2 is equal to 0.50. Each cell is averaged over 1,000 simulations.
different distribution in the template and the group of treated units, the effect estimate obtained from 1,000 matched pairs constructed using algorithm $M_{\text{opt}}$ is clearly biased from $\text{ATE}_{\text{target}}$, and the percentage of bias increases (i) as the distributions of $X_1$ in the target population and in the group of OBS treated units become increasingly dissimilar, that is, as $\hat{\beta}$ increases, and (ii) as effect modification becomes more pronounced, that is, from $\beta(X_1) = 2$ (constant) to $\beta(X_1) = 2 - 0.2X_1$ (mild) to $\beta(X_1) = 2 - X_1$ (strong). In the most adversarial scenario with $\beta(X_1) = 2 - X_1$, the percentage bias of $\hat{\beta}_{M_{\text{opt}}}$ can be as large as 25%. We need to stress that although $\hat{\beta}_{M_{\text{opt}}}$ may not be generalized to the target population, it is an internally-valid estimator for the average treatment effect on the treated.

Our proposed algorithm outperforms $M_{\text{opt}}$ in bias reduction against $\text{ATE}_{\text{target}}$ in all six different implementations under all data-generating processes considered in this simulation study, although the gain in bias reduction differs from implementation to implementation. The gain is most pronounced when (i) $k$ is small so that a smaller treated group bearing more resemblance to the target template is constructed, and (ii) $\lambda$ is small so that the matching algorithm gives priority to resemblance to the template.

3.3 Unmeasured confounding, computation cost, and comparison to the two-step procedure

In Web Appendix D.2, we designed a simulation study to investigate the role of unmeasured confounding. In the presence of unmeasured confounding in the observational database, the bias consisted of two parts, a generalization bias and an unmeasured confounding bias. The proposed algorithm helps remove most of the generalization bias so that researchers could better focus on the unmeasured confounding bias. In Web Appendix D.3, we compared the computation cost of a network-flow-based algorithm and an MIP-based algorithm implemented in the R package designmatch based on an open source optimization routine GLPK. We found that the network-flow-based method largely outperforms the MIP-based method in this comparison, although we note that the performance of MIP-based methods would improve when implemented using a more powerful commercial optimization routine like Gurobi. In Web Appendix D.4, we compared the proposed method to a two-step procedure based on solving an optimal bipartite matching twice. We found that the proposed algorithm outperformed the two-step procedure under a wide range of scenarios and that the gain was not sensitive to the choice of the tuning parameter $\lambda$.

4 REVISITING THE WOMEN’s HEALTH INITIATIVE STUDY

4.1 Background and our goal

The WHI is a combined clinical trial and observational study. Postmenopausal women were screened for clinical trial eligibility; those who were ineligible or unwilling to participate in the trial were enrolled in the observational study. The design of the WHI study is described in The Women’s Health Initiative Study Group (1998). One important goal of the WHI clinical trial is to evaluate the hypothesized cardioprotective effect of postmenopausal hormone therapy, following a substantial body of evidence from observational studies (The Women’s Health Initiative Study Group, 1998). The WHI estrogen-plus-progestin (E+P for short) trial found a rather surprising elevation in coronary heart disease risk (Writing Group for the Women’s Health Initiative Investigators, 2002) and sparked a lot of discussion regarding the discrepancy between clinical trial and observational study results.

Many authors have speculated on why findings differed. Some major concerns include: (i) potential bias due to unmeasured confounding in observational studies (Humphrey et al., 2002); (ii) biological differences between trial participants and those in the observational study including the difference in time since menopause at hormone therapy initiation (Hernán et al., 2008; Michels & Manson, 2003; Prentice et al., 2009; Willett et al., 2006), among others. Table 2 summarizes some important baseline covariates in the WHI trial and associated observational study and illustrates some of these concerns. Compared to past and never users, current users of estrogen-plus-progestin in the WHI observational study are younger, less likely to be black or Hispanic, more educated, and have more physical activity episodes per week. On the other hand, baseline covariates of participants in the control and intervention groups are similar in the WHI trial by virtue of randomization. There is also a tangible difference in socioeconomic status and smoking status between trial and observational study participants. Moreover, trial participants initiated their HRT at a much older age compared to the current users in the observational study.

Our proposed matching algorithm could help alleviate these concerns. The first concern regarding the validity of the observational study could be alleviated if matched observational study participants are balanced for a large number of baseline covariates including detailed demographics, preexisting comorbid conditions, and personal habits. The concerns about the comparability between OBS and RCT participants in their cardiovascular risk...
### TABLE 2 Important baseline characteristics of the WHI observational study and WHI trial subjects.

|                      | WHI observational data |                      | WHI trial |                      |
|----------------------|------------------------|----------------------|-----------|----------------------|
|                      | Never/past users (n = 75303) | Current Users (n = 18340) | Control n = 8102 | E+P intervention (n = 8506) |
| **Age at screening** | 64.30 (7.37)           | 60.84 (6.69)         | 63.33 (7.11)     | 63.23 (7.13)         |
| **Race/Ethnicity**  |                        |                      |                       |                       |
| White                | 61704 (81.9)           | 16285 (88.8)         | 6805 (84.0)       | 7141 (84.0)          |
| Black/Hispanic       | 10166 (13.5)           | 1074 (5.9)           | 989 (12.2)        | 1019 (12.0)          |
| Other                | 3433 (4.6)             | 981 (5.3)            | 308 (3.8)         | 346 (4.1)            |
| **Education**        |                        |                      |                       |                       |
| College or above     | 30418 (40.4)           | 10044 (54.8)         | 3011 (37.2)       | 3111 (36.6)          |
| Some college         | 28074 (37.3)           | 5853 (31.9)          | 3060 (37.8)       | 3357 (39.5)          |
| High school diploma/GED | 13068 (17.4)       | 2050 (11.2)          | 1609 (19.9)       | 1615 (19.0)          |
| Other                | 3743 (5.0)             | 393 (2.1)            | 422 (5.2)         | 423 (5.0)            |
| **Blood Pressure**   |                        |                      |                       |                       |
| Systolic             | 128 (18.12)            | 123 (16.92)          | 128 (17.53)       | 128 (17.63)          |
| Diastolic            | 75 (9.42)              | 74 (8.99)            | 76 (9.09)         | 76 (9.12)            |
| **BMI**              | 27.62 (5.98)           | 25.84 (5.15)         | 28.50 (5.91)      | 28.46 (5.82)         |
| **Smoking**          |                        |                      |                       |                       |
| NA                   | 1004 (1.3)             | 220 (1.2)            | 98 (1.2)          | 83 (1.0)             |
| Current Smoker       | 4850 (6.4)             | 940 (5.1)            | 838 (10.4)        | 880 (10.3)           |
| Never Smoked         | 38296 (50.9)           | 8710 (47.5)          | 3999 (49.4)       | 4178 (49.1)          |
| Past Smoker          | 31048 (41.3)           | 8448 (46.1)          | 3157 (39.0)       | 3362 (39.5)          |
| **No. of PA episodes**|                       |                      |                       |                       |
| Total                | 5.28 (4.12)            | 5.80 (4.12)          | 4.77 (4.06)       | 4.74 (4.10)          |
| Medium to strenuous  | 2.94 (3.37)            | 3.59 (3.57)          | 2.58 (3.24)       | 2.50 (3.21)          |
| **HRT initiation**   |                        |                      |                       |                       |
| NA                   | 67259 (89.3)           | 0                    | 6706 (82.8)       | 0                    |
| Age at initiation    | 52.04 (7.35)           | 53.87 (6.84)         | 53.57 (6.53)      | 61.93 (8.15)         |
| Previous E+P use in years | 0.47 (2.12)    | 7.02 (5.64)          | 0.64 (2.23)       | 0.70 (2.34)          |
| **Unopposed estrogen use ever** |           |                      |                       |                       |
| Yes                  | 32705 (43.4)           | 2170 (11.8)          | 865 (10.7)        | 903 (10.6)           |
| No                   | 42598 (56.6)           | 16170 (88.2)         | 7237 (89.3)       | 7603 (89.4)          |
| **Reproductive history** |                 |                      |                       |                       |
| No ovary removed     | 48303 (64.1)           | 16906 (92.2)         | 7705 (95.1)       | 8083 (95.0)          |

*Note: Mean (SE) are reported for continuous variables and count (%) for categorical variables.*

Profile, HRT initiation time, and previous estrogen-plus-progestin usage could be mitigated by treating the RCT cohort as a template and constructing matched OBS pairs that resemble this template in these aspects. Our proposed algorithm cannot fully address the difference in HRT initiation time and previous estrogen-plus-progestin usage between RCT and OBS participants: current users in the observational study had used HRT for more than 7 years on average, while participants of the trial had used HRT for less than a year on average. This difference is largely due to the design of the WHI study: the observational study enrolled a cross-section of post-menopausal women that included many who had been using the therapy for years (Willett et al., 2006), while most trial
participants were never-users of the therapy by the time of randomization. This lack of overlap makes it virtually impossible to completely balance this aspect between RCT and OBS participants; however, in our opinion, closing this important gap as much as the data allow helps evaluate Willett et al.’s (2006) arguments that (i) HRT initiation time could be an important effect modifier, an argument supported by Hernán et al.’s (2008) re-analysis of the Nurses’ Health Study; and (ii) the interaction between this effect modification and the difference in the “time since menopause” between the WHI RCT and OBS participants could potentially explain the difference in findings.

4.2 Template and three matched samples

Our goal is to construct well-matched pairs from the WHI observational study that resemble the WHI trial participants in the following sense: (i) matched pairs resemble the WHI trial participants in cardiovascular risk profile, and (ii) treated participants in the matched pairs resemble the WHI trial intervention group in their HRT initiation time and previous estrogen-plus-progestin usage as much as possible. Our template consists of a random sample of 1,000 WHI trial participants in the intervention group with the following covariates: risk factors listed in Table 2 plus the HRT initiation time.

We applied the proposed matching algorithm to constructing matched pairs of different sizes corresponding to choosing different \( k \)’s in Section 2.3. Our desired matched OBS cohort would have similar cardiovascular risk profile as the template and are closely matched for many other baseline covariates to maximally guard against the unmeasured confounding bias due to the non-randomized nature of the WHI observational study. In addition to risk factors listed in Table 2, we further matched on participants’ region, partners’ education level, income level, marital status, reproductive history, and eight important preexisting conditions.

Table 3 summarizes results from two matched samples constructed using the proposed algorithm with different parameters. Match M1 constructed 10,000 matched pairs of two observational study participants. Match M1 used a Mahalanobis distance with an estimated generalizability score caliper (caliper size = 0.05) and a Mahalanobis distance with an estimated propensity score caliper (caliper size = 0.15) on the left and right parts of the network depicted in Figure 2, respectively. Because of the large sample size (\( \approx 100,000 \) in the observational study), we applied a “hard” caliper in the sense that edges connecting one OBS treated and one OBS control units are removed whenever they differ in their estimated propensity score by more than the caliper size; in this way, the network is sparsified and computation is boosted. We set \( \lambda = 100 \) to give priority to internal validity of the matched comparison. Match M2 is similar to M1 except that we only formed 3,000 matched pairs. Lastly, match M0 formed 18,340 matched pairs exhausting every treated observational study participant. The balance table and propensity score distributions of M0 can be found in Web Appendix E.

Judging from internal validity, all three matched cohorts are acceptable: the absolute standardized mean differences (SMDs) of most cardiovascular risk factors and additional OBS covariates are less than 0.1, or one-tenth of one pooled standard deviation (Silber et al., 2001), and many are below 0.01. However, the three designs differ, sometimes significantly, in their resemblance to the trial population. In particular, design M0 differs from the trial participants in their cardiovascular risk profile (e.g., the percentage of black/Hispanic is 6% in M0 compared to 12% in the template) and HRT initiation age (54 in M0 compared to 62 in the trial population). On the other hand, both M1 and M2 are similar to the trial population in the cardiovascular risk profile: the percentage of white (black/Hispanic) women was 85% (10%) in M1, 84% (12%) in M2 and 84% (12%) in the template, compared to 89% (6%) in M0; the percentage with a college degree or above is 41% in M1, 36% in M2, and 38% in the template, compared to 55% in M0. In fact, the matched treated group in either M1 or M2 is now indistinguishable from the template based on the cardiovascular risk profile as judged by Gagnon-Bartsch and Shem-Tov’s classification permutation test (2019). Moreover, the treated group of M1 now initiated their HRT at an average age of 57 with an average 5.54 year of HRT usage prior to enrolling in the WHI study, and the treated group of M2 initiated their HRT at an average age of 61 with an average of 3.24 year previous HRT usage. The HRT initiation age in the design M2 is closer to that of the trial population, though a potentially meaningful discrepancy in the previous HRT usage persists (3.24 vs. 0.70 years).

4.3 Comparing survival outcomes

Figure 3 plots the Kaplan–Meier curves (Kaplan & Meier, 1958) in the treatment and control groups in each study design. Compared to all controls in the unadjusted analysis (top left panel of Figure 3), matched controls in M0 have higher survival probability, reflecting the fact that these matched controls are now more identical to the treated OBS participants and healthier; however, the cardioprotective effect of HRT persists (\( P \)-value = 0.020 according to O’Brien and Fleming’s (1987) matched-pair Prentice–Wilcoxon test) in M0 and is consistent with the previous analysis of the WHI observational data (Prentice et al., 2005). Compared to their counterparts in M0, treated and matched comparison groups in M1 have lower survival
### Table 3: Balance Table of Baseline Covariates in WHI Observational Study before Matching, Two Matched Samples Constructed using the Proposed Algorithm, and the Matched Template.

|                      | OBS treated | Matched treated M1 | Matched treated M2 | Template | Matched control M2 | Matched control M1 | OBS control | SMD M1 | SMD M2 |
|----------------------|-------------|--------------------|--------------------|----------|--------------------|--------------------|-------------|--------|--------|
| **Sample Size**      |             |                    |                    |          |                    |                    |             |        |        |
| \( N \)              | 18,340      | 10,000             | 3,000              | 1,000    | 3,000              | 10,000             | 75,303      |        |        |

**HRT Prior Usage**

|                      |             |                    |                    |          |                    |                    |             |        |        |
|----------------------|-------------|--------------------|--------------------|----------|--------------------|--------------------|-------------|--------|--------|
| **Age at initiation**|             |                    |                    |          |                    |                    |             |        |        |
| NA                   | 0           | 0                  | 0                  | 0        | 0.88               | 0.86               | 0.89        |        |        |
| Age                  | 53.87       | 56.83              | 60.67              | 61.93    | 53.14              | 52.44              | 52.04       |        |        |
| Previous E+P use, yrs| 7.02        | 5.54               | 3.24               | 0.70     | 0.45               | 0.51               | 0.47        |        |        |

**Covariates Collected in RCT and OBS**

|                      |             |                    |                    |          |                    |                    |             |        |        |
|----------------------|-------------|--------------------|--------------------|----------|--------------------|--------------------|-------------|--------|--------|
| **Age at screening** |             |                    |                    |          |                    |                    |             |        |        |
|                       | 60.84       | 62.21              | 63.57              | 63.10    | 63.35              | 62.15              | 64.30       | 0.01   | 0.03   |

**Race/Ethnicity**

|                      |             |                    |                    |          |                    |                    |             |        |        |
|----------------------|-------------|--------------------|--------------------|----------|--------------------|--------------------|-------------|--------|--------|
| White                | 0.89        | 0.85               | 0.84               | 0.84     | 0.84               | 0.85               | 0.82        | −0.00  | −0.01  |
| Black/Hispanic       | 0.06        | 0.10               | 0.12               | 0.12     | 0.12               | 0.10               | 0.14        | 0.00   | 0.01   |

**Education**

|                      |             |                    |                    |          |                    |                    |             |        |        |
|----------------------|-------------|--------------------|--------------------|----------|--------------------|--------------------|-------------|--------|--------|
| College or above     | 0.55        | 0.41               | 0.36               | 0.38     | 0.36               | 0.40               | 0.40        | 0.01   | 0.00   |
| Some college         | 0.32        | 0.41               | 0.43               | 0.40     | 0.43               | 0.41               | 0.37        | −0.00  | 0.01   |
| High school diploma/GED | 0.11   | 0.15               | 0.17               | 0.17     | 0.17               | 0.16               | 0.17        | −0.01  | −0.02  |

**Blood pressure**

|                      |             |                    |                    |          |                    |                    |             |        |        |
|----------------------|-------------|--------------------|--------------------|----------|--------------------|--------------------|-------------|--------|--------|
| Systolic             | 123.27      | 125.26             | 126.69             | 126.99   | 126.28             | 125.11             | 127.86      | 0.01   | 0.02   |
| Diastolic            | 74.03       | 74.42              | 75.01              | 75.66    | 74.81              | 74.36              | 74.91       | 0.01   | 0.02   |
| BMI                  | 25.85       | 27.01              | 28.22              | 28.46    | 27.61              | 26.78              | 27.61       | 0.04   | 0.11   |

**Smoking**

|                      |             |                    |                    |          |                    |                    |             |        |        |
|----------------------|-------------|--------------------|--------------------|----------|--------------------|--------------------|-------------|--------|--------|
| Current smoker       | 0.05        | 0.09               | 0.11               | 0.10     | 0.11               | 0.09               | 0.06        | 0.00   | 0.01   |
| Never smoked         | 0.47        | 0.47               | 0.49               | 0.51     | 0.47               | 0.47               | 0.51        | 0.00   | 0.04   |
| Past smoker          | 0.46        | 0.43               | 0.39               | 0.37     | 0.41               | 0.43               | 0.41        | −0.00  | −0.05  |

(Continues)
| No. of PA episodes | OBS treated | Matched treated M1 | Matched treated M2 | Template | Matched control M2 | Matched control M1 | OBS control | SMD M1 | SMD M2 |
|--------------------|-------------|-------------------|------------------|---------|------------------|------------------|-------------|--------|--------|
| Total              | 5.80        | 5.23              | 4.79             | 4.71    | 4.83             | 5.22             | 5.29        | 0.00   | −0.01  |
| Medium to strenuous| 3.58        | 3.05              | 2.59             | 2.56    | 2.67             | 3.01             | 2.94        | 0.01   | −0.02  |
| Unopposed estrogen use ever | 0.12 | 0.10 | 0.10 | 0.10 | 0.10 | 0.10 | 0.43 | 0.00 | −0.00 |
| No ovary removed   | 0.92        | 0.95              | 0.95             | 0.95    | 0.95             | 0.64             | −0.00       | 0.00   |        |

**Additional OBS Covariates**

**Region**

|        | OBS treated | Matched treated M1 | Matched treated M2 | Template | Matched control M2 | Matched control M1 | OBS control | SMD M1 | SMD M2 |
|--------|-------------|-------------------|------------------|---------|------------------|------------------|-------------|--------|--------|
| Midwest| 0.22        | 0.22              | 0.22             | 0.24    | 0.24             | 0.22             | −0.04       | −0.04  |        |
| Northeast | 0.18     | 0.20              | 0.21             | 0.24    | 0.21             | 0.24             | −0.03       | −0.07  |        |
| South  | 0.25        | 0.25              | 0.25             | 0.24    | 0.24             | 0.26             | 0.01        | 0.02   |        |

**Partner's education**

|                          | OBS treated | Matched treated M1 | Matched treated M2 | Template | Matched control M2 | Matched control M1 | OBS control | SMD M1 | SMD M2 |
|-------------------------|-------------|-------------------|------------------|---------|------------------|------------------|-------------|--------|--------|
| College or above        | 0.43        | 0.37              | 0.32             | 0.31    | 0.35             | 0.30             | 0.03        | 0.03   |        |
| Some college            | 0.16        | 0.18              | 0.18             | 0.18    | 0.18             | 0.17             | 0.00        | −0.01  |        |
| High school diploma/GED | 0.07        | 0.08              | 0.09             | 0.09    | 0.09             | 0.09             | −0.02       | −0.01  |        |

**Income**

|                          | OBS treated | Matched treated M1 | Matched treated M2 | Template | Matched control M2 | Matched control M1 | OBS control | SMD M1 | SMD M2 |
|-------------------------|-------------|-------------------|------------------|---------|------------------|------------------|-------------|--------|--------|
| Below 35K               | 0.23        | 0.29              | 0.35             | 0.37    | 0.30             | 0.40             | −0.03       | −0.03  |        |
| 35K - 75K               | 0.42        | 0.43              | 0.41             | 0.39    | 0.43             | 0.36             | 0.00        | 0.03   |        |
| Above 75K               | 0.29        | 0.22              | 0.18             | 0.16    | 0.20             | 0.16             | 0.04        | 0.04   |        |

**Marital status**

|                          | OBS treated | Matched treated M1 | Matched treated M2 | Template | Matched control M2 | Matched control M1 | OBS control | SMD M1 | SMD M2 |
|-------------------------|-------------|-------------------|------------------|---------|------------------|------------------|-------------|--------|--------|
| Married                 | 0.69        | 0.67              | 0.62             | 0.62    | 0.66             | 0.60             | 0.02        | 0.02   |        |
| Divorced/Widowed        | 0.26        | 0.29              | 0.33             | 0.33    | 0.30             | 0.35             | −0.01       | 0.00   |        |

**Employment status**

|                          | OBS treated | Matched treated M1 | Matched treated M2 | Template | Matched control M2 | Matched control M1 | OBS control | SMD M1 | SMD M2 |
|-------------------------|-------------|-------------------|------------------|---------|------------------|------------------|-------------|--------|--------|
| Yes                     | 0.45        | 0.40              | 0.36             | 0.35    | 0.40             | 0.32             | 0.00        | 0.01   |        |
| No                      | 0.53        | 0.58              | 0.62             | 0.63    | 0.58             | 0.65             | −0.00       | −0.01  |        |

**Reproductive history**

|                          | OBS treated | Matched treated M1 | Matched treated M2 | Template | Matched control M2 | Matched control M1 | OBS control | SMD M1 | SMD M2 |
|-------------------------|-------------|-------------------|------------------|---------|------------------|------------------|-------------|--------|--------|
| Oral contraceptive use ever | 0.53     | 0.48              | 0.43             | 0.42    | 0.47             | 0.37             | 0.00        | 0.01   |        |
| OC duration in years    | 5.59        | 5.55              | 5.47             | 5.18    | 5.32             | 5.16             | 0.06        | 0.08   | (Continues) |
TABLE 3 (Continued)

| Preexisting Conditions | OBS treated | Matched treated M1 | Matched treated M2 | Template | Matched control M2 | Matched control M1 | OBS control | SMD M1 | SMD M2 |
|------------------------|-------------|---------------------|-------------------|----------|-------------------|-------------------|-------------|--------|--------|
| Stroke                 | 0.01        | 0.01                | 0.01              |          | 0.01              | 0.01              | 0.01        |        | −0.00  |
| MI                     | 0.01        | 0.02                | 0.02              |          | 0.02              | 0.02              | 0.03        | −0.01  | 0.01   |
| CHF                    | 0.00        | 0.01                | 0.01              |          | 0.00              | 0.00              | 0.01        | 0.02   | 0.01   |
| Liver diseases         | 0.02        | 0.02                | 0.02              |          | 0.02              | 0.02              | 0.02        | −0.01  | 0.03   |
| Hypertension           | 0.25        | 0.29                | 0.34              |          | 0.31              | 0.28              | 0.35        | 0.02   | 0.07   |
| Fracture               | 0.10        | 0.12                | 0.16              |          | 0.13              | 0.12              | 0.14        | 0.03   | 0.09   |
| CABG/PTCA              | 0.01        | 0.01                | 0.02              |          | 0.01              | 0.02              | 0.02        | −0.01  | 0.04   |
| BRCA                   | 0.01        | 0.01                | 0.01              |          | 0.01              | 0.01              | 0.07        | 0.00   | 0.00   |

Note: Match M1 constructed 10,000 matched pairs and Match M2 3,000 matched pairs.

5. SUMMARY

We proposed a statistical matching algorithm that constructs well-matched pairs from observational databases that resemble a target population. By designing databases that resemble a target population, one can potentially (i) better reconcile the sometimes conflicting study findings, and (ii) answer a clinical/epidemiological query for a scientifically meaningful population. In a typical observational study, researchers use matching to adjust for confounders associated with the treatment assignment and clinical outcomes. With an added complication of mimicking a target population, key covariates to be matched to the template could potentially consist of a subset of confounders that are suspected to be effect modifiers, because it is the interaction between effect modifiers that introduces the generalization bias (Stuart et al., 2011). If the overlap between the treated and control groups and between the treated and the template could afford, recommending adjusting for any suspected confounders could seem appropriate.

We introduced the generalization bias that is attributed to effect modifiers’ distributions that differ from the target population. If we adjust for covariates that are suspected to be effect modifiers, we may be able to reduce the generalization bias. In this way, matching allows one to adjust for confounders associated with the treatment assignment and clinical outcomes, in a way that resembles the target population. By designing databases that resemble a target population, one can potentially (i) better reconcile the sometimes conflicting study findings, and (ii) answer a clinical/epidemiological query for a scientifically meaningful population. In a typical observational study, researchers use matching to adjust for confounders associated with the treatment assignment and clinical outcomes. With an added complication of mimicking a target population, key covariates to be matched to the template could potentially consist of a subset of confounders that are suspected to be effect modifiers, because it is the interaction between effect modifiers that introduces the generalization bias (Stuart et al., 2011). If the overlap between the treated and control groups and between the treated and the template could afford, recommending adjusting for any suspected confounders could seem appropriate.

We introduced a statistical matching algorithm that constructs well-matched pairs from large observational databases that resemble a target population. By designing databases that resemble a target population, one can potentially (i) better reconcile the sometimes conflicting study findings, and (ii) answer a clinical/epidemiological query for a scientifically meaningful population. In a typical observational study, researchers use matching to adjust for confounders associated with the treatment assignment and clinical outcomes. With an added complication of mimicking a target population, key covariates to be matched to the template could potentially consist of a subset of confounders that are suspected to be effect modifiers, because it is the interaction between effect modifiers that introduces the generalization bias (Stuart et al., 2011). If the overlap between the treated and control groups and between the treated and the template could afford, recommending adjusting for any suspected confounders could seem appropriate.

We introduced a statistical matching algorithm that constructs well-matched pairs from large observational databases that resemble a target population. By designing databases that resemble a target population, one can potentially (i) better reconcile the sometimes conflicting study findings, and (ii) answer a clinical/epidemiological query for a scientifically meaningful population. In a typical observational study, researchers use matching to adjust for confounders associated with the treatment assignment and clinical outcomes. With an added complication of mimicking a target population, key covariates to be matched to the template could potentially consist of a subset of confounders that are suspected to be effect modifiers, because it is the interaction between effect modifiers that introduces the generalization bias (Stuart et al., 2011). If the overlap between the treated and control groups and between the treated and the template could afford, recommending adjusting for any suspected confounders could seem appropriate.

We introduced a statistical matching algorithm that constructs well-matched pairs from large observational databases that resemble a target population. By designing databases that resemble a target population, one can potentially (i) better reconcile the sometimes conflicting study findings, and (ii) answer a clinical/epidemiological query for a scientifically meaningful population. In a typical observational study, researchers use matching to adjust for confounders associated with the treatment assignment and clinical outcomes. With an added complication of mimicking a target population, key covariates to be matched to the template could potentially consist of a subset of confounders that are suspected to be effect modifiers, because it is the interaction between effect modifiers that introduces the generalization bias (Stuart et al., 2011). If the overlap between the treated and control groups and between the treated and the template could afford, recommending adjusting for any suspected confounders could seem appropriate.
A key application of the proposed method is the thriving field of "target trial emulation" using data from observational databases. As discussed near the end of Section 1.2, the RCT is still the gold standard for drawing causal conclusions. Well-balanced matched pairs constructed from observational studies may still suffer from hidden bias. As Rosenbaum (2017, Chapter 6) put it: "the absence of an obvious reason to think that two groups are different falls well short of a compelling reason to think they are the same." This is not saying that evidence from observational studies should be completely dismissed; carefully designed observational studies could be valuable in closing important knowledge gaps.

We applied the proposed method to directly address some hypotheses raised in Willett et al. (2006) regarding the discrepancy between the WHI observational study and clinical trial findings on HRT’s effect on coronary heart diseases. We found that a matched cohort study constructed from the WHI observational data still supported a cardioprotective effect of HRT, and this cardioprotective effect persisted even after the designed matched samples were forced to resemble the trial population in their cardiovascular risk profile. However, in a matched design that further mimics the trial intervention group in the HRT initiation age, the cardioprotective effect seemed to disappear. Our findings provide some evidence for Willett et al.’s (2006) argument that HRT initiation age might have played an important role in explaining the observational study and trial discrepancy.

ACKNOWLEDGMENTS
The author would like to thank the co-editor, associate editor and two reviewers for their dedication and patience in the review process, which has made this final version a much improved one.

DATA AVAILABILITY STATEMENT
The data that support the findings of this paper are available from the National Heart, Lung, and Blood Institute. Restrictions apply to the availability of these data, which were used under license for this paper. Data are available from the author with the permission of the NHLBI.
REFERENCES

Bennett, M., Vielma, J.P. & Zubizarreta, J.R. (2020) Building representative matched samples with multi-valued treatments in large observational studies. *Journal of Computational and Graphical Statistics*, 29, 744–757.

Chen, H. & Small, D.S. (2022) New multivariate tests for assessing covariate balance in matched observational studies. *Biometrics*, 78, 202–213.

Dahabreh, I.J., Robertson, S.E., Tchetgen, E.J., Stuart, E.A. & Hernán, M.A. (2019) Generalizing causal inferences from individuals in randomized trials to all trial-eligible individuals. *Biometrics*, 75, 685–694.

Franklin, J.M., Patorno, E., Desai, R.J., Glynn, R.J., Martin, D., Quinto, K., Pawar, A., Bessette, L.G., Lee, H., Garry, E.M., Gautam, N. & Schneeweiss, S. (2021) Emulating randomized clinical trials with nonrandomized real-world evidence studies: first results from the RCT DUPLICATE initiative. *Circulation*, 143, 1002–1013.

Franklin, J.M., Pawar, A., Martin, D., Glynn, R.J., Levenson, M., Temple, R. & Schneeweiss, S. (2020) Nonrandomized real-world evidence to support regulatory decision making: process for a randomized trial replication project. *Clinical Pharmacology & Therapeutics*, 107, 817–826.

Gagnon-Bartsch, J. & Shem-Tov, Y. (2019) The classification permutation test: a flexible approach to testing for covariate imbalance in observational studies. *Annals of Applied Statistics*, 13, 1464–1483.

Hernán, M.A., Alonso, A., Logan, R., et al. (2008) Observational studies analyzed like randomized experiments: an application to postmenopausal hormone therapy and coronary heart disease. *Epidemiology* (Cambridge, Mass.), 19, 766.

Ho, D.E., Imai, K., King, G. & Stuart, E.A. (2007) Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political Analysis*, 15, 199–236.

Humphrey, L.L., Chan, B.K. & Sox, H.C. (2002) Postmenopausal hormone replacement therapy and the primary prevention of cardiovascular disease. *Annals of Internal Medicine*, 137, 273–284.

Kaplan, E.L. & Meier, P. (1958) Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53, 457–481.

Karmakar, B., Small, D.S. & Rosenbaum, P.R. (2019) Using approximation algorithms to build evidence factors and related designs for observational studies. *Journal of Computational and Graphical Statistics*, 28, 698–709.

Korte, B.H. & Vygen, J. (2011) *Combinatorial Optimization*, vol. 1. Springer.

MacKay, E.J., Zhang, B., Heng, S., et al. (2021) Association between transesophageal echocardiography and clinical outcomes after coronary artery bypass graft surgery. *Journal of the American Society of Echocardiography*, 34, 571–581.

Metkus, T.S., Thibault, D., Grant, M.C., et al. (2021) Transesophageal echocardiography in patients undergoing coronary artery bypass graft surgery. *Journal of the American College of Cardiology*, 78, 112–122.

Michels, K.B. & Manson, J.E. (2003) Postmenopausal hormone therapy. *Circulation*, 107, 1830–1833.

Nattino, G., Lu, B., Shi, J., Lemeshow, S. & Xiang, H. (2021) Triplet matching for estimating causal effects with three treatment arms: a comparative study of mortality by trauma center level. *Journal of the American Statistical Association*, 116, 44–53.

O’Brien, P.C. & Fleming, T.R. (1987) A paired Prentice–Wilcoxon test for censored paired data. *Biometrics*, 43, 169–180.

Prentice, R.L., Langer, R., Stefanick, M.L., Howard, B.V., Pettinger, M., Anderson, G., Barad, D., Curb, J.D., Kotchen, J., Kuller, L., Limacher, M., Wactawski-Wende, J. & Women’s Health Initiative Investigators, (2005) Combined postmenopausal hormone therapy and cardiovascular disease: toward resolving the discrepancy between observational studies and the women’s health initiative clinical trial. *American Journal of Epidemiology*, 162, 404–414.

Prentice, R.L., Manson, J.E., Langer, R.D., Anderson, G.L., Pettinger, M., Jackson, R.D., Johnson, K.C., Kuller, L.H., Lane, D.S., Wactawski-Wende, J., Bzryska, R., Allison, M., Ockene, J., Sarto, G. & Rossouw, J.E. (2009) Benefits and risks of postmenopausal hormone therapy when it is initiated soon after menopause. *American Journal of Epidemiology*, 170, 12–23.

Rosenbaum, P. (2017) *Observation and experiment: an introduction to causal inference*. Harvard University Press.

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

Rosenbaum, P.R. (2002) *Observational studies*. Springer.

Rosenbaum, P.R. (2010) *Design of observational studies*. Springer.

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

Rosenbaum, P.R. & Rubin, D.B. (1985) Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician*, 39, 33–38.

Rubin, D.B. (1979) Using multivariate matched sampling and regression adjustment to control bias in observational studies. *Journal of the American Statistical Association*, 74, 318–328.

Silber, J.H., Rosenbaum, P.R., Ross, R.N., Ludwig, J.M., Wang, W., Niknam, B.A., Mukherjee, N., Saynisch, P.A., Even-Shoshan, O., Kelz, R.R. & Fleisher, L.A. (2014) Template matching for auditing hospital cost and quality. *Health Services Research*, 49, 1446–1474.

Silber, J.H., Rosenbaum, P.R., Trudeau, M.E., Even-Shoshan, O., Chen, W., Zhang, X. & Mosher, R.E. (2001) Multivariate matching and bias reduction in the surgical outcomes study. *Medical Care*, 39, 1048–1064.

Stuart, E.A. (2010) Matching methods for causal inference: a review and a look forward. *Statistical Science*, 25, 1–21.

Stuart, E.A., Cole, S.R., Bradshaw, C.P. & Leaf, P.J. (2011) The use of propensity scores to assess the generalizability of results from randomized trials. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 174, 369–386.

The Women’s Health Initiative Study Group, (1998) Design of the Women’s Health Initiative clinical trial and observational study. *Controlled Clinical Trials*, 19, 61–109.

Willett, W.C., Manson, J.E., Grodstein, F., Stampfer, M.J. & Colditz, G.A. (2006) Re: “combined postmenopausal hormone therapy and cardiovascular disease: toward resolving the discrepancy between observational studies and the women’s health initiative clinical trial”. *American Journal of Epidemiology*, 163, 1067–1068.
Writing Group for the Women’s Health Initiative Investigators, (2002) Risks and benefits of estrogen plus progestin in healthy postmenopausal women: principal results from the women’s health initiative randomized controlled trial. *JAMA*, 288, 321–333.

Yu, R. (2021) Evaluating and improving a matched comparison of antidepressants and bone density. *Biometrics*, 77, 1276–1288.

Zhang, B., Small, D., Lasater, K., McHugh, M., Silber, J. & Rosenbaum, P. (2021) Matching one sample according to two criteria in observational studies. *Journal of the American Statistical Association*, 118, 1–34.

Zubizarreta, J.R. (2012) Using mixed integer programming for matching in an observational study of kidney failure after surgery. *Journal of the American Statistical Association*, 107, 1360–1371.

**SUPPORTING INFORMATION**

Web Appendices referenced in Sections 1–4 are available with this paper at the Biometrics website on Wiley Online Library. R package match2C implements the proposed method. Code necessary to reproduce simulation results can be found in the supplement.

**How to cite this article:** Zhang, B. (2023) Efficient algorithms for building representative matched pairs with enhanced generalizability. *Biometrics*, 79, 3981–3997.

[https://doi.org/10.1111/biom.13919](https://doi.org/10.1111/biom.13919)