Towards better reconciling randomized controlled trial and observational study findings: Efficient algorithms for building representative matched samples with enhanced external validity

Bo Zhang *

Vaccine and Infectious Disease Division, Fred Hutchinson Cancer Research Center, Seattle, WA, U.S.A.

Abstract: Many recent efforts center on assessing the ability of real-world evidence (RWE) generated from nonrandomized observational data to provide results that are compatible with those from randomized controlled trials (RCTs). One noticeable endeavor is the RCT DUPLICATE initiative (Franklin et al., 2020). To better reconcile findings from observational and trial data, it is desirable to eliminate differences between the RCT and corresponding observational study populations. We outline an efficient, network-flow-based statistical matching algorithm that designs well-matched pairs from observational data that mimic the covariates’ distribution of a target population, e.g., the RCT study population or a 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 study populations’ cardiovascular risk profile, but seemed to disappear in a study design that further adjusted for the HRT initiation age and previous estrogen-plus-progestin use. The proposed method is integrated into the R package match2C.

Keywords: generalizability, matching, RCT DUPLICATE Initiative, Women’s Health Initiative

*Bo Zhang is an incoming assistant professor of biostatistics at the Fred Hutchinson Cancer Research Center. Email: bozhan950119@gmail.com
1 Introduction

1.1 RCT DUPLICATE Initiative: Comparing Observational Studies to Randomized Controlled Trials (RCTs)

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 10 randomized controlled trials (RCTs) investigating very similar clinical questions. The RCT DUPLICATE initiative aims to build an empirical evidence base for real world data through large-scale replication of RCTs and understand to which clinical questions and by what analytic tools researchers could draw credible causal conclusions from retrospective, non-randomized data (e.g., electronic health records, administrative claims databases, and diseases registries).

To enable a better 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 study 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:

[I]nclusion 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 preexisting comorbid conditions between RCT and observational study populations are common in Franklin et al. (2021)’s emulation studies. For instance, the Saxagliptin and Cardiovascular Outcomes in Patients with Type 2 Diabetes Mellitus (SAVOR-TIMI 53) study (Scirica et al., 2013), one of the ten trials Franklin et al. (2021) emulated, enrolled 33.1% female and 37.8% with history of myocardial infarction (MI); the observational study emulating the SAVOR-TIMI 53 trial, however, consisted of 46.8% female and only 11.2% with history of MI (Franklin et al., 2021, Table 1). These differences persisted in the final matched samples constructed by Franklin et al. (2021), and could partially explain the disagreement in effect estimates derived from RCTs and observational studies.

The ongoing RCT emulation study led by the RCT DUPLICATE initiative and many similar endeavors to better reconcile RCT and observational study findings (e.g., the Women’s Health Ini-
tiative Study; see, e.g., Prentice et al. (2005) and Hernán et al. (2008)) motivate us to develop a transparent, easy-to-use algorithm that designs matched samples homogeneous in many covariates while mimicking a target population (e.g., the RCT population) in some key covariates (e.g., potential effect modifiers), so that the scientific conclusions drawn from the observational data could be better generalized to other context.

1.2 A Naive Method, An Existing Method, and A New Approach

One naive approach to creating homogeneous matched pairs resembling a target template would first select a subset of treated units from the observational data via matching on covariates collected by both the RCT and the observational study. In the second step, control units in the observational data are then matched to the treated units selected by the first step. This strategy suffers from a major drawback. Take the RCT DUPLICATE initiative as an example. Although the benchmark SAVOR-TIMI 53 trial reported only around 20 baseline covariates, Franklin et al. (2021) matched on more than 120 to guard against unmeasured confounding in their emulation with observational data. In the first step, there are many ways to design a smaller treated group similar to the RCT population in 20 RCT covariates; however, it is difficult to determine which treated group designed from the first step should be utilized to form the eventual match in the second step. A selected treated group similar to the RCT population in 20 RCT covariates could have poor overlap with the control units in the observational database in the other more than 100 covariates; see Figure 1 for a toy example that illustrates this phenomenon. In many cases, we prefer well-matched treated and control groups to first ensure internal validity of the matched cohort study; similarity to some target population should be taken as a perk. The naive approach precludes a principled trade-off between matched cohort study’s internal validity and its generalizability to a target population.

Recently, Bennett et al. (2020) developed a mixed-integer-programming-based (MIP-based) approach that builds representative matched samples. MIP-based matching algorithms encode requirements for different versions of covariate balance as constraints in a mathematical program. Although recent advancements in computing power have made it more practical to solve large-scale MIP problems with many complex constraints, the general MIP problems are still theoretically intractable or NP-hard. From a very practical perspective, MIP-based methods require installing a powerful commercial optimization routine (e.g., Gurobi or IBM CPLEX), which is proprietary and could be an obstacle to some researchers. This being said, if one finds no difficulty installing the solver and that the match can be readily delivered
Figure 1: A toy example illustrating the limitation of a naive two-step approach. Suppose our goal is to create one treated-to-control pair. In the first step, treated units are matched to the template to obtain a candidate treated group. In this toy example, both $T_1$ and $T_2$ are good match to the template judging from two RCT covariates $X_1$ and $X_2$. However, in view of the reservoir of control units, $T_1$ is a better pick as no control unit has an $X_3$ value similar to that of $T_2$.

by MIP-based methods, then this is great and one should do that. Other approaches to adjusting for sample selection bias include weighting and doubly-robust methods (see, e.g., Cole and Stuart, 2010; Stuart et al., 2011a; Dahabreh et al., 2019); see Colnet et al. (2020); Degtiar and Rose (2021) for a more comprehensive review.

Compared to MIP-based statistical matching methods, network-flow-based methods (see Supplemental Material A for a literature review) 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; Austin, 2011a; Rassen et al., 2012). The primary goal of this article is to outline an efficient, network-flow-based algorithm that designs matched pairs from observational data with close resemblance to a target population. We demonstrate the usefulness of the proposed method by revisiting the Women’s Health Initiative (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 persisted in a design that adjusted for the cardiovascular risk profile differences between the observational and trial data, but seemed to disappear in a design that further adjusted for the HRT initiation age and previous estrogen-plus-progestin use, resonating with similar findings in the seminal work by Hernán et al. (2008).
2 Methodology

2.1 Basic Network Structure: Vertices and Edges

We describe and illustrate the proposed method using the RCT emulation study, although the idea can be immediately generalized to accommodate an arbitrary target population other than an RCT population. We refer to study units from the randomized controlled trial as RCT units and those from the observational database as OBS units for short.

Figure 2 depicts a basic version of the proposed network structure. There are \( R \) treated units from the target RCT. These RCT units \( \{\kappa_1, \kappa_2, \ldots, \kappa_R\} \) are represented by nodes labeled \( \kappa_r, r = 1, 2, \ldots, R \). There are \( T \geq R \) OBS treated units and \( C \geq T \) OBS control units from some administrative database. OBS treated units \( \{\tau_1, \tau_2, \ldots, \tau_T\} \) are represented twice in the network, by nodes labeled \( \tau_t \) and \( \bar{\tau}_t \), \( t = 1, 2, \ldots, T \), and OBS control units \( \{\gamma_1, \gamma_2, \ldots, \gamma_C\} \) are represented by nodes labeled \( \gamma_c \), \( c = 1, 2, \ldots, C \). In addition to \( R + 2 \times T + C \) nodes representing RCT and OBS study units, there is a source node \( \xi \) and a sink node \( \bar{\xi} \), so that the network consists of

\[
\mathcal{V} = \{\xi, \kappa_1, \ldots, \kappa_R, \tau_1, \ldots, \tau_T, \bar{\tau}_1, \ldots, \bar{\tau}_T, \gamma_1, \ldots, \gamma_C, \xi, \bar{\xi}\}.
\]

(1)

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:

\[
\mathcal{E} = \{(\xi, \kappa_r), (\kappa_r, \tau_t), (\tau_t, \bar{\tau}_t), (\bar{\tau}_t, \gamma_c), (\gamma_c, \xi), \ r = 1, \ldots, R, \ t = 1, \ldots, T, \ c = 1, \ldots, C\}.
\]

(2)

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

2.2 Basic Network Structure: Capacity and Matched Samples

To transform a statistical matching problem into an appropriate network-flow optimization problem, one needs to carefully design the cost and capacity of each edge. Fix an integer \( k \) such that \( 1 \leq k \leq \lfloor T/R \rfloor \), and consider constructing \( N = kR \) treated-to-control matched pairs from the observational database. For instance, the SAVOR-TIMI 53 trial consists of \( R = 8,280 \) units assigned saxagliptin, the intervention under evaluation, and three retrospective databases available for emulation (Optum, MarketScan, and Medicare) consist of \( T = 91,082 \) subjects exposed to
saxagliptin according to [Franklin et al. (2021)](https://www.clinicaltrials.gov/ct2/show/NCT03936023)’s study protocol registered at ClinicalTrials.gov (identifier NCT03936023). Researcher could in principle choose any integer $k$ between 1 and $\left\lfloor \frac{91,082}{8,280} \right\rfloor = 11$ 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, $\text{cap}((\xi, \kappa_r)) = k$ for $r = 1, \ldots, R$, and all other edges have capacity 1. In Figure 2, the source $\xi$ supplies $kR$ units of flow, the sink $\bar{\xi}$ absorbs $kR$ units of flow, while all other nodes preserve the flow by simply passing them along [Ahuja et al. 1988; Bertsekas 1991; 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 $\{0, 1, 2, \ldots, k\}$ such that (i) all capacity constraints are respected, i.e., $0 \leq f(e) \leq \text{cap}(e)$, $e \in \mathcal{E}$, (ii) $kR$ units of flow are supplied at $\xi$ and absorbed at $\bar{\xi}$, i.e., $\sum_{r=1}^{R} f((\xi, \kappa_r)) = kR$ and $\sum_{c=1}^{C} f((\gamma_c, \bar{\xi})) = kR$, and (iii) the flow is preserved at all nodes other than $\xi$ and $\bar{\xi}$, i.e., $\sum_{(a,b) \in \mathcal{E} \setminus \{\xi, \bar{\xi}\}} f((a,b)) = \sum_{(b,c) \in \mathcal{E} \setminus \{\xi, \bar{\xi}\}} f((b,c))$ for all $b \in V \setminus \{\xi, \bar{\xi}\}$.

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 3 correspond to one (out of $\binom{4}{3} \times 6 \times 5 \times 4 = 480$) 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 outputs matched samples.
Formally, the matched samples $\mathcal{M}$ is defined by

$$\mathcal{M} = \{(\tau_t, \gamma_c) \text{ such that } f((\tau_t, \tau_t)) = f((\tau_t, \gamma_c)) = 1\}. \quad (3)$$

For instance, the matched samples returned by the feasible flow in Figure 3 consist of $\mathcal{M} = \{(\tau_1, \gamma_1), (\tau_2, \gamma_5), (\tau_3, \gamma_4)\}$.

Figure 3: A feasible flow $f(\cdot) : \mathcal{E} \mapsto \{0, 1\}$ in a dense network. Thick, black lines correspond to edges with $f(e) = 1$ while light gray lines correspond to edges with $f(e) = 0$. The feasible flow yields a matched sample of three pairs: $\mathcal{M} = \{(\tau_1, \gamma_1), (\tau_2, \gamma_5), (\tau_3, \gamma_4)\}$.

2.3 Basic Network Structure: Cost, Probability of Participation, Propensity Score

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

Costs associated with edges $(\kappa_r, \tau_t)$ play an important role in forcing selected OBS samples to mimic RCT units in some covariates. Suppose that each RCT unit $\kappa_r$ is associated with a vector of covariates $\tilde{x}$ and each OBS unit $(\tilde{x}, x)$. As in the RCT DUPLICATE example, $\tilde{x}$ contains roughly 20 covariates that both RCT and observational database collected while $x$ consist of more than 100 additional covariates available only in the claims database. Let $\delta_{\kappa_r, \tau_t}(\tilde{x})$ denote a measure of covariate distance between $\kappa_r$ and $\tau_t$ in $\tilde{x}$. The cost associated with an edge of the form $(\kappa_r, \tau_t)$ is
equal to \(\delta_{\kappa_r,\tau_t}(\overline{x})\), i.e.,

\[
\text{cost}\{(\kappa_r, \tau_t)\} = \delta_{\kappa_r,\tau_t}(\overline{x}).
\] (4)

One intriguing strategy equals \(\delta_{\kappa_r,\tau_t}(\overline{x})\) to a scalar “balancing score” of \(\overline{x}\) (Rosenbaum and Rubin, 1983), so that matching on this balancing score stochastically balances \(\overline{x}\). In the context of generalizing RCT’s effect estimates to a target population, Stuart et al. (2011a) defined and studied the conditional probability of selecting into the RCT group rather than the OBS group, which is referred to as the “probability of participation.” In practice, researchers could collect the RCT treated units \(K\) and OBS treated units \(T\), and estimate the conditional probability of being selected into the RCT given covariates \(\overline{x}\) using a logistic regression (Stuart et al., 2011a).

Lastly, let \(\Delta_{\tau_t,\gamma_c}\{(\overline{x}, x)\}\) denote a measure of distance between OBS treated unit \(\tau_t\) and OBS control unit \(\gamma_c\) in their observed covariates \((\overline{x}, x)\). The cost associated with an edge \((\tau_t, \gamma_c)\) is set to \(\Delta_{\tau_t,\gamma_c}\{(\overline{x}, x)\}\), i.e.,

\[
\text{cost}\{\{\tau_t, \gamma_c\}\} = \Delta_{\tau_t,\gamma_c}\{(\overline{x}, x)\}.
\] (5)

Different specifications of \(\Delta_{\tau_t,\gamma_c}\{(\overline{x}, x)\}\) have been extensively studied in the literature (Rosenbaum, 2002, 2010; Stuart, 2010). Perhaps the most widely used strategy is to equal \(\Delta_{\tau_t,\gamma_c}\{(\overline{x}, x)\}\) to the absolute difference in Rosenbaum and Rubin (1983)’s propensity score. Alternatively, one may let \(\Delta_{\tau_t,\gamma_c}\{(\overline{x}, x)\}\) represent the Mahalanobis distance in \((\overline{x}, x)\) (Cochran and Rubin, 1973; Rubin, 1979), or the Mahalanobis distance within a propensity score caliper (Rosenbaum and Rubin, 1985).

### 2.4 Minimum Cost Flow, Complexity, Trade-Off Between Internal and External Validity

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

\[
\text{cost}(f) = \sum_{e \in \mathcal{E}} f(e) \cdot \text{cost}(e) = \sum_{(\kappa_r, \tau_t) \in \mathcal{E} : f((\kappa_r, \tau_t)) = 1} \delta_{\kappa_r,\tau_t}(\overline{x}) + \sum_{(\tau_t, \gamma_c) \in \mathcal{E} : f((\tau_t, \gamma_c)) = 1} \Delta_{\tau_t,\gamma_c}\{(\overline{x}, x)\},
\] (6)

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 \(O(|\mathcal{V}| \cdot |\mathcal{E}| + |\mathcal{V}|^2 \log(|\mathcal{V}|))\) operations (Korte and Vygen, 2011) and in the proposed network, we have \(|\mathcal{V}| = O(C), |\mathcal{E}| = O(C^2)\), so that the computation complexity \(O(|\mathcal{V}| \cdot |\mathcal{E}| + |\mathcal{V}|^2 \log(|\mathcal{V}|))\) simplifies to \(O(C^3)\). We discuss ways to sparsify the network and speed up computation in Supplemental Material B. In the statistical computing software R, the minimum cost flow can be found via Bertsekas (1991)’s auction algorithm.
implemented by Bertsekas and Tseng (1988) and made available by Hansen (2007) and Pimentel et al. (2015).

There is a tension between internal and external validity in a matched observational study. Take the RCT DUPLICATE initiative as an example. In their emulation of the SAVOR-TIMI 53 trial using observational data, Franklin et al. (2021) matched closely on and balanced almost 120 observed covariates; however, the effect estimate obtained from observational data (HR: 0.81; 95% CI: (0.76, 0.86)) was significantly different from the RCT effect estimate (HR: 1.00; 95% CI: (0.89, 1.12)). There are many possible explanations in this large discrepancy, one of which is a quite large difference in baseline characteristics of the RCT versus OBS populations, including a disparity in some important preexisting conditions like MI. It is one possibility that both the observational study and RCT effect estimates are internally valid; they are just reporting effects for different populations.

According to our formulation, the cost \( S_1 = \sum \left( \tau_{t, \gamma_c} \right) \in \mathcal{E} \quad \Delta_{\tau_{t, \gamma_c}} \{ (\bar{x}, x) \} \) in expression (6) measures the homogeneity between the matched treated and control groups. Well-matched samples have a small \( S_1 \) value and better internal validity. On the other hand, the cost \( S_2 = \sum \left( \kappa_r, \tau_t \right) \in \mathcal{E} \quad \delta_{\kappa_r, \tau_t} (\bar{x}) \) measures how much the matched samples mimic a target population, and a small \( S_2 \) corresponds to improved generalizability to the target population. To enable researchers explore the trade-off between matched samples’ internal and external validity, we replace \( \Delta_{\tau_{t, \gamma_c}} \{ (\bar{x}, x) \} \) with \( \lambda \cdot \Delta_{\tau_{t, \gamma_c}} \{ (\bar{x}, x) \} \) for some \( \lambda > 0 \) so that \( \text{cost}(f) \) becomes a weighted average of \( S_1 \) and \( S_2 \):

\[
\text{cost}(f) = \sum_{e \in \mathcal{E}} f(e) \cdot \text{cost}(e) = \lambda \cdot \left\{ \sum_{\left( \kappa_r, \tau_t \right) \in \mathcal{E} : f\{\left( \kappa_r, \tau_t \right) \} = 1} \delta_{\kappa_r, \tau_t} (\bar{x}) \right\} + \lambda \cdot \left\{ \sum_{\left( \tau_{t, \gamma_c} \right) \in \mathcal{E} : f\{\left( \tau_{t, \gamma_c} \right) \} = 1} \Delta_{\tau_{t, \gamma_c}} \{ (\bar{x}, x) \} \right\}. \quad (7)
\]

A small \( \lambda \) value gives priority to matched samples’ internal validity, while a large \( \lambda \) value generalizability to the target population. A similar weighting scheme is also used in Zhang et al. (2021) but for a different purpose.

### 3 Simulation Study and Software Availability

We illustrated the proposed method with an example in Supplemental Material C, and described a comprehensive simulation study in Supplemental Material D. In the simulation study, we examined how the matched samples delivered by the proposed method improves the performance of the
downstream outcome analysis. In particular, we considered settings where the estimand of interest is the sample average treatment effect of a target template, the treatment effect is heterogeneous, and the effect modifiers’ distributions are different in the target and observational data. Our proposed algorithm is able to achieve significant bias reduction compared to statistical matching algorithms that pays no attention to the target template. The proposed matching algorithm is integrated in the package match2C available via the statistical computing software R (R Core Team, 2021) with a detailed tutorial.

4 Revisiting the Women’s Health Initiative (WHI) Study

4.1 Background and Our Goal

The Women’s Health Initiative (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 concerning the discrepancy between clinical trial and observational study results.

Many authors have speculated on why results would differ dramatically between the trial and observational study. Some major concerns include: (i) potential bias due to unmeasured confounding in observational studies (Humphrey et al., 2002; Rutter, 2007; Yu et al., 2021); (ii) biological differences between trial participants and those in the observational study (Michels and Manson, 2003); (iii) differences in time since menopause at hormone therapy initiation (Prentice et al., 2005; Willett et al., 2006; Rossouw et al., 2007; Hernán et al., 2008; Prentice et al., 2009). Table 1 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, have slightly smaller BMI and 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.

Table 1: Important baseline characteristics of the WHI observational study and WHI trial subjects. Mean (SE) are reported for continuous variables and count (%) for categorical variables.

|fac| WHI Observational Data | | | | WHI Trial | | |
|---|------------------------|---|---|---|---|---|---|
| | Never/Past Users | | | | Control | | |
| | (n = 75303) | | | | (n = 8102) | | |
| | Current Users | | | | E+P Intervention | | |
| | (n = 18340) | | | | (n = 8506) | | |
| Age at screening | 64.30 (7.37) | | | 63.33 (7.11) | | |
| | 60.84 (6.69) | | | 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) | | 10453 (59.8) | | 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) |

Our proposed matching algorithm is suited for addressing concerns (i), (ii), and (iii). The first concern of the internal validity of the observational study can be greatly alleviated if matched OBS samples are balanced for a large number of baseline covariates including detailed demographics, preexisting comorbid conditions, personal habits, etc. Rich covariate information collected by the WHI study offered us an opportunity to do this. The second and third concerns about the comparability between OBS and RCT samples in their cardiovascular risk profile and HRT initiation time can be at least partially addressed by treating the RCT samples as a template and forcing
matched OBS samples to resemble the RCT cohort in these aspects.

Reconciling RCT and OBS results goes beyond addressing these three concerns. One concern that cannot be adequately addressed is the difference in estrogen-plus-progestin usage time prior to the WHI study between RCT and OBS samples: 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. The lack of overlap between RCT and OBS samples in previous estrogen-plus-progestin usage (due to the study design) makes it virtually impossible to balance this aspect between RCT and OBS samples using any matching tool; nevertheless, in our opinion, addressing or just alleviating some major concerns (e.g., concerns (i), (ii), and (iii)) helps researchers critically examine the remaining explanations of the discrepancy and is a meaningful step towards ultimately reconciling the RCT and OBS findings.

4.2 Template and Three Matched Samples

We aim to construct well-matched samples from the WHI observational study that resemble the WHI trial units in the following sense: (i) matched treated and control units resemble the WHI trial units in cardiovascular risk profile, and (ii) matched treated units resemble the WHI trial intervention group in the HRT initiation time.

To this end, our template consists of a random sample of 1000 WHI trial units in the intervention group with the following covariates: risk factors listed in Table 1 plus the HRT initiation time. In many applications, researchers may not have individual-level data but only the joint distribution of important baseline covariates; see, e.g., Cole and Stuart (2010) when they try to generalize the AIDS Clinical Trial Group 320 Study (ACTG 320 Trial) to the HIV-infected individuals in the United States. If this is the case, researchers could readily simulate individual-level data subject to the marginal joint distribution constraints, as suggested by Bennett et al. (2020).

We applied the proposed matching algorithm to constructing matched samples of different sizes corresponding to choosing different $k$’s in Section 2.2. Our desired matched OBS samples would have similar cardiovascular risk profile as the template, and 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 particular, in addition to risk factors listed in Table
we further matched on subjects’ region, partner’s education level, income level, marital status, reproductive history, and eight important preexisting conditions.

Table 2 summarizes results from two matched samples constructed using the proposed template matching algorithm with different parameters. Match M1 constructed 10,000 matched pairs of two units: one OBS treated unit and the other OBS control unit. Match M1 used a Mahalanobis distance with an estimated generalizability score caliper (caliper size = 0.05) on the left and a Mahalanobis distance with an estimated propensity score caliper (caliper size = 0.15) on the right; see Figure 3. Because the large sample size (≈ 100,000 in the observational study), we applied a “hard” caliper in the sense that edges connecting OBS treated and OBS control units are removed whenever the two units differ in their estimated propensity score by more than the caliper size; in this way, the network structure is largely sparsified and computation is boosted. We set λ = 10 to give some priority to mimicking the target template. Match M2 is similar to M1 except that we only formed 3,000 matched pairs. Lastly, match M0 formed 18,340 matched pairs exhausting each and every treated OBS unit. The balance table and distributions of the estimated propensity score in the treated and matched control groups of M0 can be found in the Supplemental Material E.

Judging from internal validity, all three matched samples M0, M1, and M2 are acceptable; in fact, the absolute standardized mean differences (SMDs) of all cardiovascular risk factors and additional OBS covariates are less than 0.1, or one-tenth of one pooled standard deviation (Silber et al., 2001; Austin and Stuart, 2015); in fact, most absolute SMDs are much less than 0.1. However, the three matches differ, sometimes significantly, in their resemblance to the target template. In particular, matched samples M0 differ from the target template in its cardiovascular risk profile (e.g., the percentage of black/Hispanic is 6% in M0 compared to 12% in the template; the percentage with a college degree or above is 55% in M0 compared to merely 38% in the template), and HRT initiation age (54 in M0 compared to 62 in the template, which translates to a 6.30-year difference in the previous HRT usage).

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% (11%) 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 40% in M1, 36% in M2, and 38% in the template, compared to 55% in M0. Moreover, the treated group of M1 now initiated their HRT at an average age of 57 with an average 5.61-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 3.28-year previous HRT usage. The HRT initiation age of treated units in M2 is now very close to that of the trial population, with a difference less than one year, though there is still a potentially meaningful discrepancy in the previous HRT usage (3.28 years in M2 vs 0.70 years in the template).

| Study Samples | Internal Validity | Cardiovascular Risk Profile | HRT Initiation | Previous E+P Use | Cardioprotective Effect of HRT |
|---------------|-------------------|-----------------------------|----------------|-----------------|-------------------------------|
| Before        | Very Poor         | Poor                        | Very Poor      | Very Poor       | Very Large                    |
| M0            | Good              | Poor                        | Very Poor      | Very Poor       | Persisted                     |
| M1            | Good              | Good                        | Poor           | Poor            | Persisted                     |
| M2            | Good              | Good                        | Good           | Fair            | Disappeared                   |

Table 2: A qualitative summary of the internal and external validity of observational study data before matching and three matched samples.

Table 2 gives a qualitative summary of three different study designs (M0, M1, and M2): in short, while all three designs have stellar internal validity, M2 most resembles the target template, with M1 coming in the second, and M0 the last.

4.3 Comparing Survival Outcomes

Figure 4 plots the Kaplan-Meier curves (Kaplan and Meier, 1958) in the treatment and control groups in each study design. Compared to the entire reservoir of control OBS units in the unadjusted analysis, matched control units in M0 have much higher survival probabilities, reflecting the fact that these matched controls are more identical to the treated units and a lot healthier; however, the cardioprotective effect of HRT still exists ($P$-value = 0.02 according to O’Brien and Fleming (1987)’s matched-pair Prentice-Wilcoxon test) in M0 and is consistent with the previous analysis of the WHI observational data (Prentice et al., 2005, 2006). Compared to their counterparts in M0, both treated and matched comparison groups in M1 have lower survival probabilities; however, a qualitatively similar cardioprotective effect persists in the design M1 ($P$-value = 0.09). Recall that M1 has largely mimicked the cardiovascular risk profile of the target template; therefore, it seems that the difference in cardiovascular risk profile between observational study and trial populations is not sufficient in explaining the inconsistency in the trial and observational study results. Compared to M0 and M1, M2 best resembles the trial population in their HRT initiation age and previous HRT use; indeed, when we examine the survival outcomes of matched OBS units in M2, the
cardioprotective effect seemed to disappear ($P$-value = 0.65). Three comparisons facilitated by three different study designs M0, M1, and M2 seemed to support Willett et al. (2006)’s assessment that HRT initiation age and previous HRT usage played a key role in the discrepancy between WHI observational study and trial findings, and resonates with similar findings in Hernán et al. (2008).

![Figure 4: Survival outcomes for coronary heart disease (CHD) in the unadjusted observational study samples (top left), design M0 (top right), design M1 (bottom left), and design M2 (bottom right).](image)

5 Summary

Motivated by the ongoing work undertaken by the RCT DUPLICATE Initiative, we proposed a statistical matching algorithm that constructs well-matched samples from large observational databases while mimicking a target population like the RCT population. By designing matched samples that resemble a target population, empirical researchers could potentially (i) better reconcile the sometimes conflicting RCT and observational study findings, and (ii) answer a clinical/epi-
demiological query for a scientifically meaningful population.

We applied the proposed method to investigate the discrepancy between the Women’s Health Initiative (WHI) observational study and clinical trial findings on HRT’s effect on coronary heart diseases. The method facilitated some interesting findings. In particular, 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 mimic the WHI trial population in the cardiovascular risk profile. However, after we further designed matched samples resembling the WHI trial intervention group in the HRT initiation age, the cardioprotective effect seemed to disappear. Our findings seemed to provide some evidence for the argument that HRT initiation age might have played a major role in explaining the observational study and trial discrepancy (Willett et al., 2006; Hernán et al., 2008).
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. Match M1 constructed 10000 matched pairs and Match M2 3000 matched pairs.

| Sample Size | OBS Treated | Matched Treated M1 | Matched Treated M2 | Template | Matched Control M2 | Matched Control M1 | OBS Control | SMD M1 | SMD M2 |
|-------------|-------------|--------------------|--------------------|----------|--------------------|--------------------|-------------|--------|--------|
| N           | 18340       | 10000              | 3000               | 1000     | 3000               | 10000              | 75303       |        |        |

**HRT Prior Usage**

| Age at initiation | NA | 0 | 0 | 0 | 0 | 0.88 | 0.87 | 0.89 |
|-------------------|----|---|---|---|---|------|------|------|
| Age               | 53.87 | 57.02 | 60.96 | 61.93 | 52.95 | 52.21 | 52.04 |
| Previous E+P use, yrs | 7.02 | 5.61 | 3.28 | 0.70 | 0.50 | 0.51 | 0.47 |

**Covariates Collected in RCT and OBS**

| Blood pressure | Systolic | 123.27 | 125.65 | 127.23 | 126.99 | 125.43 | 127.71 | 127.86 | 0.01 | -0.03 |
|----------------|----------|--------|--------|--------|--------|--------|--------|--------|------|-------|
| Diastolic      | 74.03    | 74.47  | 75.00  | 75.66  | 74.89  | 74.57  | 74.91  | 0.01   | 0.01  |
| BMI            | 25.85    | 27.10  | 28.19  | 28.46  | 27.70  | 26.78  | 27.61  | 0.06   | 0.09  |

**Smoking**

| Current smoker | 0.05 | 0.09 | 0.11 | 0.10 | 0.11 | 0.09 | 0.06 | 0.01 | 0.00 |
|----------------|------|------|------|------|------|------|------|------|------|
| Never smoked   | 0.47 | 0.47 | 0.48 | 0.51 | 0.47 | 0.47 | 0.51 | 0.00 | 0.03 |
| Past smoker    | 0.46 | 0.42 | 0.40 | 0.37 | 0.41 | 0.43 | 0.41 | -0.01 | -0.03 |

**No. of PA episodes**

| Total | 5.80 | 5.25 | 4.77 | 4.71 | 4.79 | 5.33 | 5.29 | -0.02 | -0.00 |
|-------|------|------|------|------|------|------|------|-------|-------|
| Medium to strenuous | 3.58 | 3.05 | 2.61 | 2.56 | 2.57 | 3.07 | 2.94 | -0.00 | 0.01 |

**Additional OBS Covariates**

| Region | Midwest | 0.22 | 0.22 | 0.23 | 0.23 | 0.22 | 0.22 | -0.01 | 0.00 |
|--------|---------|------|------|------|------|------|------|-------|------|
|        | Northeast | 0.18 | 0.19 | 0.20 | 0.20 | 0.21 | 0.04 | -0.01 | -0.06 |
|        | South    | 0.25 | 0.25 | 0.25 | 0.25 | 0.26 | 0.26 | -0.02 | -0.00 |
|        | College or above | 0.43 | 0.36 | 0.31 | 0.30 | 0.34 | 0.30 | 0.03 | 0.03 |
|        | Some college | 0.16 | 0.19 | 0.18 | 0.18 | 0.18 | 0.17 | 0.00 | 0.00 |
|        | High school diploma/GED | 0.07 | 0.08 | 0.09 | 0.09 | 0.08 | 0.09 | -0.00 | 0.01 |
| Income: | Below 35K | 0.23 | 0.29 | 0.35 | 0.37 | 0.31 | 0.40 | -0.03 | -0.04 |
|         | 35K - 75K | 0.42 | 0.43 | 0.42 | 0.39 | 0.42 | 0.36 | 0.01 | 0.06 |
|         | Above 75K | 0.29 | 0.21 | 0.17 | 0.17 | 0.20 | 0.16 | 0.03 | -0.00 |
| Marital status: | Married | 0.69 | 0.66 | 0.62 | 0.61 | 0.65 | 0.60 | 0.03 | 0.02 |
|          | Divorced/Widowed | 0.26 | 0.29 | 0.33 | 0.35 | 0.31 | 0.35 | -0.03 | -0.04 |
| Employment status: | Yes | 0.45 | 0.39 | 0.34 | 0.34 | 0.39 | 0.32 | 0.00 | 0.00 |
|          | No | 0.53 | 0.59 | 0.64 | 0.64 | 0.59 | 0.65 | 0.00 | 0.00 |
| Reproductive history: | No ovary removed | 0.92 | 0.91 | 0.92 | 0.92 | 0.91 | 0.64 | 0.00 | 0.01 |
|          | Oral contraceptive use ever | 0.53 | 0.46 | 0.41 | 0.41 | 0.46 | 0.37 | 0.00 | 0.01 |
|          | OC duration in years | 5.59 | 5.52 | 5.43 | 5.15 | 5.27 | 5.16 | 0.07 | 0.08 |
| Preexisting Conditions: | Stroke | 0.01 | 0.01 | 0.01 | 0.01 | 0.02 | 0.01 | 0.00 | 0.00 |
|          | MI | 0.01 | 0.02 | 0.02 | 0.02 | 0.02 | 0.02 | 0.02 | 0.05 |
|          | CHF | 0.00 | 0.01 | 0.01 | 0.01 | 0.01 | 0.01 | 0.01 | 0.01 |
|          | Liver diseases | 0.02 | 0.02 | 0.02 | 0.03 | 0.02 | 0.02 | -0.03 | -0.02 |
|          | Hypertension | 0.25 | 0.29 | 0.34 | 0.33 | 0.28 | 0.35 | 0.03 | 0.02 |
|          | Fracture | 0.10 | 0.13 | 0.15 | 0.14 | 0.13 | 0.14 | -0.00 | 0.04 |
|          | CABG/PTCA | 0.01 | 0.02 | 0.02 | 0.02 | 0.02 | 0.02 | 0.02 | 0.02 |
|          | BRCA | 0.01 | 0.01 | 0.01 | 0.01 | 0.01 | 0.07 | 0.00 | 0.00 |
Online Supplemental Materials for “Towards better reconciling randomized controlled trial and observational study findings: Efficient algorithms for building representative matched samples with enhanced external validity”

Supplemental Material A: Statistical Matching, Network Flow, Recent Advancement

Matching and subclassification are widely used in empirical research to embed noisy, non-randomized observational data into an approximate randomized controlled experiment, and facilitate analyzing data as such [Ho et al. 2007; Rubin 2007; Rubin et al. 2008; Rosenbaum 2002, 2010, 2020; Stuart 2010; Rassen et al. 2012; Imbens 2015]. 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. Prior to Rosenbaum (1989), statisticians and practitioners leveraged a “greedy heuristic” in pair matching that starts with finding a matched pair with a minimum pre-specified distance, removes the pair from further consideration, and iterates the process until finding one control unit for each treated one in the study cohort. Some often-used distance metrics include the Mahalanobis distance in metric-based matching (Cochran and Rubin 1973; Rubin 1980), absolute difference in the estimated propensity score in propensity-score matching (Rosenbaum and Rubin 1983), or a combination of both (Rosenbaum and Rubin 1985).

Rosenbaum (1989)’s key insight is to represent units to be matched, treated and control, as vertices in a bipartite graph and recast the statistical problem of constructing a matched control group as finding a minimum-cost network flow, a standard combinatorial optimization problem for which very efficient algorithms exist. Figure 1 illustrates this paradigm using a small example of three treated and five control units. Readers should imagine that the source node on the far left emits three units of flow, one unit 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 $j = 1, 2, ..., 5$, each have capacity 1 and are associated with a cost $\text{cost}\{e(T_i, C_j)\}$ equal to some covariate distance between $T_i$ and $C_j$. These three units of flow arriving at nodes $T_1$, $T_2$, and $T_3$ would choose to flow in such a way that minimizes the total cost; for instance, Figure 1 displays a minimum-cost flow (bold, black lines) consisting of $\{e(T_1, C_3), e(T_2, C_1), e(T_3, C_4)\}$. In this way, three matched pairs $\{(T_1, C_3), (T_2, C_1), (T_3, C_4)\}$ are constructed. The network structure displayed in Figure 1 can be readily modified to accommodate
many useful design aspects, including exact matching, 1-to-\( k \) matching, full matching (Hansen, 2004; Hansen and Klopfer, 2006), subset matching (Rosenbaum et al., 2007), among others, and is the cornerstone of some most influential statistical matching packages like optmatch (Hansen, 2007) and MatchIt (Stuart et al., 2011b) in the statistical computing software R (R Core Team, 2013). Researchers have also proposed network-sparsification techniques to further facilitate statistical matching on large administrative datasets consisting of hundreds of thousands of observations (Pimentel et al., 2015; Yu et al., 2020).

More recently, Zhang et al. (2021) introduced a minimum-cost network flow algorithm built around a “tripartite graph,” where treated units appear twice, on the far left and far right, with control units sandwiched between them. Efforts to balance high-dimensional covariates (e.g., via the stochastic balancing property of the propensity score or by directly minimizing the earthmover distance between marginal distributions) are represented on the right, while efforts to find close pairings (e.g., exact matching on potential effect modifiers or quantiles of Hansen (2008)’s prognostic score) are represented on the left. By separating two sometimes-conflicting objectives, pairing and balancing, the network built around a tripartite graph is shown to deliver matched samples that are homogeneous in a few key covariates while maintaining good balance in many other covariates (Zhang et al., 2021).

![Network-flow representation of a small matching problem. Five control units \( C_1, C_2, \ldots, C_5 \) are to be matched to three treated units \( T_1, T_2, \) and \( T_3 \). Each edge connecting \( T_i, i = 1, 2, 3, \) and \( C_j, j = 1, 2, 3, 4, 5, \) is associated with a flow capacity (equal to 1 in pair matching) and a cost equal to a pre-specified covariate distance between unit \( T_i \) and unit \( C_j \).](image)

Representing statistical matching as a two-part network as depicted in Figure 1 is an attractive, general conceptual framework. Recent advancements in network-flow-based matching algorithms often involve adding an additional “category” layer to the two-part network and using this additional layer to balance one or more nominal covariates in the treated and matched control groups.
More recently, Zhang et al. (2021) introduced a minimum-cost network flow algorithm built around a “tripartite graph,” where treated units appear twice, on the far left and far right, with control units sandwiched between them. Efforts to balance high-dimensional covariates (e.g., via the stochastic balancing property of the propensity score or by directly minimizing the earthmover’s distance between marginal distributions) are represented on the right, while efforts to find close pairings (e.g., exact matching on potential effect modifiers or quantiles of Hansen (2008)’s prognostic score) are represented on the left. By separating two sometimes-conflicting objectives, pairing and balancing, the network built around a tripartite graph is shown to deliver matched samples that are homogeneous in a few key covariates while maintaining good balance in many other covariates (Zhang et al., 2021).

Supplemental Material B: Additional Design and Computation Considerations

Many additional design techniques can be used in conjunction with the proposed method by modifying aspects of the basic network structure in Figure 2. For instance, it is often helpful to match exactly on the potential effect modifiers in the design stage of an observational study, so that researchers could perform subgroup analysis by doing hypothesis testing in each stratum defined by the effect modifiers; see, e.g., Lee et al. (2018a,b, 2021). If such a design is desired, then researchers should set $\Delta_{t, c} \{ (\widetilde{x}, x) \} = \infty$, or equivalently remove the edge connecting $\tau_t$ and $\gamma_c$, for all $\tau_t$ and $\gamma_c$ disagreeing on the effect modifiers. Rosenbaum (1989, Section 3.2)’s fine balance strategy that forces equi-distribution of one nominal variable in the treated and matched control groups can also be readily accommodated by including an additional “category” layer in the network; see also Rosenbaum et al. (2007); Yang et al. (2012); Pimentel et al. (2015); Zhang et al. (2021). In some practical situations, certain treated OBS units $\tau_t \in T_{sub}$ are expected to be included in the matched samples. To achieve this, keep cost $\{ (\tau_t, \tau_t) \} = 0$ for $\tau_t \in T_{sub}$ and set cost $\{ (\tau_t, \tau_{t}) \}$ to a large penalty for $\tau_t \in T \setminus T_{sub}$.

While the RCT is typically of a smaller, fixed sample size, modern observational databases could easily go beyond 1 million observations. Consider a scenario with fixed RCT sample size $R$ and growing observational data sample sizes such that $T \to \infty$, $C \to \infty$, and $T/C \to \epsilon \in (0, 1)$. Recall the computation complexity of finding a minimum cost flow is $O(|V| \cdot |E| + |V|^2 \log(|V|))$. If the network can be made sparser so that $|E| = O(|V| \log(|V|)) = O(C \log(C))$, then the complexity can be reduced from $O(C^3)$ to $O(C^2 \log(C))$. One simplest way to sparsify the network is to leverage the propensity score caliper (Rosenbaum and Rubin, 1985; Austin, 2011b) and for each $\tau_t$, keep
only edges of the form $e(\tau_t, \gamma_c)$ for $C_1$ control units closest to $\tau_t$ in the estimated propensity score value. In this way, $|\mathcal{E}| = O(C_1 \cdot C)$, and the network is sparsified and computation complexity reduced provided $C_1 = O(\log C)$. Such a strategy is discussed in more detail in [Yu et al. (2020)].

**Supplemental Material C: A Small Illustrative Example**

To showcase the utility of the proposed method, we consider a simple example where we generated a target RCT of size $R = 100$, and an observational study with $T = 500$ and $C = 1500$. The RCT collects $d_1 = 5$ covariates while the observational study collects an additional 5 covariates. The RCT samples $\mathcal{K}$, OBS treated samples $\mathcal{T}$, and OBS control samples $\mathcal{C}$ have the same $\text{Normal}(0, 1)$ distribution for all but the first covariate $X_1$: $X_1 \sim \text{Normal}(0, 25, 1)$ in the RCT samples, $X_1 \sim \text{Normal}(1, 1)$ in the OBS treated samples, and $X_1 \sim \text{Normal}(0, 1)$ in the OBS control samples.

We considered three matched samples. In the first matched comparison, 500 controls were matched to 500 OBS treated units without heed of the target RCT covariate distributions. We matched this sample according to two criteria: (i) the earthmover’s distance between the distributions of the estimated propensity score in the treated group and matched control group is minimized, and (ii) subject to (i), the total robust Mahalanobis distances between treated and control units are minimized (Zhang et al., 2021). We refer to this match $M_1$. In the second matched comparison, we leveraged the proposed new network structure in Figure 2 with $k = 1$ and $\lambda = 1$ and formed 100 treated-to-control matched pairs. In this second match, we specified $\delta_{\kappa_t, \tau_t}(\overline{\mathbf{x}})$ as the absolute difference in the estimated generalizability score between $\kappa_t$ and $\tau_t$ in 5 RCT covariates, and let $\Delta_{\tau_t, \gamma_c}\{(\overline{x}, x)\}$ be the robust Mahalanobis distance within a 0.05 estimated propensity score caliper between $\tau_t$ and $\gamma_c$ in all 10 covariates. We refer to this match $M_2$. The third match $M_3$ used the same network structure and distance specifications as $M_2$, but with $k = 2$ so that 200 treated-to-control pairs were formed.

Figure 2 displays the covariate distributions of three matched samples. Before matching, the treated and control groups are vastly different in $X_1$ but otherwise similar. Match $M_1$ does a good job balancing $X_1$ via balancing the estimated propensity score; see the third boxplot in the left panel of Figure 2. However, despite within-matched-samples close resemblance, the treated group $\mathcal{T}$ and matched controls produced by $M_1$ differ systematically from the target template in the distribution of $X_1$. The 100 matched pairs returned by $M_2$ not only have good within-matched-sample-homogeneity, but also are much more similar to the target template in covariates’ distributions. The same pattern holds for the 200 matched pairs returned by $M_3$.  

4
Supplemental Material D: Simulation Studies

D.1: Goal, Structure, Measurement of Success

Our primary goal in this section is to examine how the study design delivered by different matching algorithms affects the performance of downstream statistical inference. In particular, we are interested in the case where the treatment effect is heterogeneous and the distributions of effect modifiers vary in the observational database and the target template. We generated a template \( \mathcal{K} \) with \( |\mathcal{K}| = 300 \) and \( \mathbf{X} \sim \text{Multivariate Normal}(\mu, \Sigma) \), with \( \mu = (0.25, 0, 0, 0, 0)^T \) and \( \Sigma = I_{5 \times 5} \). We consider an observational database with \( |\mathcal{T}| = 1000 \) treated and \( |\mathcal{C}| = 3000 \) 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 between template, treated \( (Z = 1) \), and control \( (Z = 0) \) units, \( \beta \): \( \mathbf{X} \sim \text{Multivariate Normal}(\mu, \Sigma) \), with \( \mu = (\theta Z, 0, 0, 0, 0)^T \) and \( \Sigma = I_{d \times d} \). We consider \( \theta = 0.25, 0.50, \) and 1.

**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 between the distributions of the estimated propensity score in the treated and matched control groups, and (ii) subject to (i), minimizing the within-matched-pair robust Mahalanobis distances. Algorithm \( \mathcal{M}_{\text{opt}} \) produces 1000 matched
pairs, and pays no attention to the target template.

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

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

**Factor 4:** Tuning parameter in $M_{\text{template}, k}$, $\lambda$: 100, 1, 0.01.

Factor 1 and 2 define the data-generating process for units in the observational database. Factor 3 and 4 define a total of $1 + 2 \times 3 = 7$ matching algorithms under consideration. As discussed in Section 2.4, the tuning parameter $\lambda$ controls the trade-off between internal and external validity of effect estimates obtained from matched samples.

For each unit, we further generate two potential outcomes according to the following data-generating process:

\[ Y(0) \sim N(0, 1), \quad Y(1) = Y(0) + \beta(X_1), \quad (8) \]

and the observed outcome satisfies $Y = Z \cdot Y(1) + (1 - Z) \cdot Y(0)$. The last factor specifies the treatment effect:

**Factor 5:** Treatment effect, $\beta(X_1)$: a constant treatment effect $\beta(X_1) = 2$, a mildly heterogeneous treatment effect $\beta(X_1) = 2 - 0.2X_1$, and a strongly heterogeneous treatment effect $\beta(X_1) = 2 - X_1$.

When the treatment effect is heterogeneous and the effect modifier $X_1$ has a different distribution in the RCT group and the OBS treated group, then the average treated effect on the treated (ATT) estimate obtained from the matched observational study cannot be automatically generalized to the RCT cohort (Stuart et al., 2011a).

There are multiple ways to analyze matched pair data. Examples include parametric t-test, randomization inference (Rosenbaum, 2002, 2010), and regression adjustment (Rubin, 1979; Ho et al., 2007). In this simulation study, we report a simple difference-in-means estimator for matched data produced by each of the 7 algorithms, and compare these 7 effect estimates to the treatment effect averaged over the target template, i.e., SATE$_{\text{target}}$. We report the average bias and mean squared error of each of the 7 effect estimates against SATE$_{\text{target}}$. 

6
D.2: Simulation Results

Table 1 summarizes the percentage of bias with respect to $\text{SATE}_{\text{target}}$ of each difference-in-means estimator $\hat{\theta}_M$ obtained from matched samples constructed from algorithm $\mathcal{M}$.

We have observed a few trends consistent with both theory and intuition. First, when the treatment is constant, there is no generalizability issue and the bias is minuscule 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 different distribution in the template and the treated units, effect estimate obtained from 1000 matched pairs constructed using algorithm $\mathcal{M}_{\text{opt}}$ is clearly biased from $\text{SATE}_{\text{target}}$, and the percentage of bias increases (i) as the distributions of $X_1$ become increasingly dissimilar in the target template and in the OBS treated units (i.e., as $b$ increases), and (ii) as effect modification becomes more dramatic (i.e., from constant to mild to strong). In the most adversarial setting considered in this simulation study, i.e., when $b = 1$ and $\beta(X_1) = 2 - X_1$, the percentage bias of $\hat{\theta}_{\mathcal{M}_{\text{opt}}}$ can be as large as 45%. We need to stress that although $\hat{\theta}_{\mathcal{M}_{\text{opt}}}$ may not be generalized to the target template, it is a perfectly internally-valid estimator for the treatment effect averaged over covariates’ distributions of OBS treated units.

Our proposed template matching algorithms outperform $\mathcal{M}_{\text{opt}}$ in bias reduction against $\text{SATE}_{\text{target}}$ in all 6 different implementations under all data-generating processes considered in this simulation study, although the gain in bias reduction differs from implementation to implementation. In particular, we observe that 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 large so that the matching algorithm gives more priority to the left part compared to the right part of the network depicted in Figure 2 i.e., resemblance to the target template in covariates’ distributions is emphasized over resemblance between matched treated and matched control groups. Both parameters, $k$ and $\lambda$, effectively allow a trade-off between the internal validity of a matched observational study and its generalizability to a target population. 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, e.g., [Franklin et al., 2020, 2021]. Provided that no outcome data are viewed, researchers can feel free to perform multiple statistical matching and select the one achieving the best trade-off between internal and external validity in an real world problem.
**Table 1:** Percentage of bias with respect to SATE\(_{\text{target}}\) of 7 difference-in-means estimator constructed from matched samples obtained from each of the 7 matching algorithms. Each cell is averaged over 1000 simulations.

| Heterogeneity Level | \(b\) | \(M_{\text{opt}}\) | \(M_{\text{template}}\) |
|---------------------|-------|-----------------|------------------|
|                     |       | \(M = 1\)       | \(M = 1\)       |
|                     |       | \(M = 10\)      | \(M = 10\)      |
|                     |       | \(k = 1\)       | \(k = 2\)       |
|                     |       | \(\lambda = 0.01\) | \(\lambda = 0.01\) |}

| Heterogeneity Level | \(b\) | \(M_{\text{opt}}\) | \(M_{\text{template}}\) |
|---------------------|-------|-----------------|------------------|
|                     |       | \(M = 1\)       | \(M = 1\)       |
|                     |       | \(M = 10\)      | \(M = 10\)      |
|                     |       | \(k = 1\)       | \(k = 2\)       |
|                     |       | \(\lambda = 0.01\) | \(\lambda = 0.01\) |}

| Heterogeneity Level | \(b\) | \(M_{\text{opt}}\) | \(M_{\text{template}}\) |
|---------------------|-------|-----------------|------------------|
|                     |       | \(M = 1\)       | \(M = 1\)       |
|                     |       | \(M = 10\)      | \(M = 10\)      |
|                     |       | \(k = 1\)       | \(k = 2\)       |
|                     |       | \(\lambda = 0.01\) | \(\lambda = 0.01\) |}

| Heterogeneity Level | \(b\) | \(M_{\text{opt}}\) | \(M_{\text{template}}\) |
|---------------------|-------|-----------------|------------------|
|                     |       | \(M = 1\)       | \(M = 1\)       |
|                     |       | \(M = 10\)      | \(M = 10\)      |
|                     |       | \(k = 1\)       | \(k = 2\)       |
|                     |       | \(\lambda = 0.01\) | \(\lambda = 0.01\) |}

| Heterogeneity Level | \(b\) | \(M_{\text{opt}}\) | \(M_{\text{template}}\) |
|---------------------|-------|-----------------|------------------|
|                     |       | \(M = 1\)       | \(M = 1\)       |
|                     |       | \(M = 10\)      | \(M = 10\)      |
|                     |       | \(k = 1\)       | \(k = 2\)       |
|                     |       | \(\lambda = 0.01\) | \(\lambda = 0.01\) |}

| Heterogeneity Level | \(b\) | \(M_{\text{opt}}\) | \(M_{\text{template}}\) |
|---------------------|-------|-----------------|------------------|
|                     |       | \(M = 1\)       | \(M = 1\)       |
|                     |       | \(M = 10\)      | \(M = 10\)      |
|                     |       | \(k = 1\)       | \(k = 2\)       |
|                     |       | \(\lambda = 0.01\) | \(\lambda = 0.01\) |
Supplemental Material E: More Details on the WHI Case Study

E.1: Plot of the Estimated Propensity Score

eFigure 3: Distributions of the estimated propensity score in the entire treated OBS group (blue), the entire control OBS group (red), and the matched control group (green).
### E.2: Balance table of M0

**eTable 2: Balance table before matching and of 18,340 matched pairs in the match M0**

|                     | OBS Treated (n = 18340) | OBS Control (n = 75303) | SMD Before | Matched Control (n=18340) | SMD M0 |
|---------------------|-------------------------|-------------------------|------------|----------------------------|--------|
| **HRT Prior Usage** |                         |                         |            |                            |        |
| Age at initiation   |                         |                         |            |                            |        |
| NA                  | 0                       | 0.89                    | 0.85       |                            |        |
| Age                 | 53.87                   | 52.04                   | 51.75      |                            |        |
| Previous E+P use, yrs | 7.02                   | 0.47                    | 0.55       |                            |        |
| **Covariates Collected in RCT and OBS** |                         |                         |            |                            |        |
| Age at screening    | 60.84                   | 64.30                   | -0.35      | 61.10                      | -0.03  |
| Race/Ethnicity      |                         |                         |            |                            |        |
| White               | 0.89                    | 0.82                    | 0.14       | 0.89                       | -0.00  |
| Black/Hispanic      | 0.06                    | 0.14                    | -0.18      | 0.06                       | 0.00   |
| Education           |                         |                         |            |                            |        |
| College or above    | 0.55                    | 0.40                    | 0.21       | 0.55                       | 0.00   |
| Some college        | 0.32                    | 0.37                    | -0.08      | 0.32                       | 0.00   |
| High school diploma/GED | 0.11                   | 0.17                    | -0.13      | 0.11                       | -0.00  |
| **Blood pressure**  |                         |                         |            |                            |        |
| Systolic            | 123.27                  | 127.86                  | -0.19      | 123.54                     | -0.01  |
| Diastolic           | 74.03                   | 74.91                   | -0.07      | 74.23                      | -0.02  |
| BMI                 | 25.85                   | 27.61                   | -0.22      | 25.99                      | -0.02  |
| Smoking             |                         |                         |            |                            |        |
| Current smoker      | 0.05                    | 0.06                    | -0.04      | 0.05                       | 0.00   |
| Never smoked        | 0.47                    | 0.51                    | -0.05      | 0.48                       | -0.00  |
| Past smoker         | 0.46                    | 0.41                    | 0.07       | 0.46                       | 0.00   |
| No. of PA episodes  |                         |                         |            |                            |        |
| Total               | 5.80                    | 5.29                    | 0.09       | 5.74                       | 0.01   |
| Medium to strenuous | 3.58                    | 2.94                    | 0.13       | 3.50                       | 0.02   |
| **Additional OBS Covariates** |                         |                         |            |                            |        |
| Region              |                         |                         |            |                            |        |
| Midwest             | 0.22                    | 0.22                    | 0.00       | 0.23                       | -0.01  |
| Northeast           | 0.18                    | 0.24                    | -0.09      | 0.19                       | -0.02  |
| South               | 0.25                    | 0.26                    | -0.02      | 0.25                       | 0.00   |
| Partner's education |                         |                         |            |                            |        |
| College or above    | 0.43                    | 0.30                    | 0.20       | 0.42                       | 0.02   |
| Some college        | 0.16                    | 0.17                    | -0.01      | 0.17                       | -0.01  |
| High school diploma/GED | 0.07                   | 0.09                    | -0.05      | 0.07                       | -0.00  |
| Income              |                         |                         |            |                            |        |
| Below 35K           | 0.23                    | 0.40                    | -0.26      | 0.24                       | -0.01  |
| 35K - 75K           | 0.42                    | 0.36                    | 0.09       | 0.44                       | -0.02  |
| Above 75K           | 0.29                    | 0.16                    | 0.22       | 0.27                       | 0.04   |
| Marital status      |                         |                         |            |                            |        |
| Married             | 0.69                    | 0.60                    | 0.13       | 0.69                       | 0.01   |
| Divorced/Widowed    | 0.26                    | 0.35                    | -0.13      | 0.27                       | -0.01  |
| Employment status   |                         |                         |            |                            |        |
| Yes                 | 0.45                    | 0.32                    | 0.19       | 0.45                       | -0.00  |
| No                  | 0.53                    | 0.65                    | -0.17      | 0.53                       | 0.00   |
| Reproductive history|                         |                         |            |                            |        |
| No ovary removed    | 0.92                    | 0.64                    | 0.51       | 0.92                       | 0.00   |
| Oral contraceptive use ever | 0.53                   | 0.37                    | 0.24       | 0.53                       | 0.00   |
| OC duration in years | 5.59                   | 5.16                    | 0.09       | 5.41                       | 0.04   |
| Preexisting Conditions |                       |                         |            |                            |        |
| Stroke              | 0.01                    | 0.02                    | -0.08      | 0.00                       | 0.01   |
| MI                  | 0.01                    | 0.03                    | -0.07      | 0.01                       | 0.00   |
| CHF                 | 0.00                    | 0.01                    | -0.06      | 0.00                       | 0.00   |
| Liver diseases      | 0.02                    | 0.02                    | -0.02      | 0.02                       | -0.01  |
| Hypertension        | 0.25                    | 0.35                    | -0.16      | 0.25                       | -0.00  |
| Fracture            | 0.10                    | 0.14                    | -0.09      | 0.11                       | -0.01  |
| CABG/PTCA           | 0.01                    | 0.02                    | -0.05      | 0.01                       | 0.00   |
| BRCA                | 0.01                    | 0.07                    | -0.22      | 0.01                       | 0.00   |
References

Ahuja, R. K., Magnanti, T. L., and Orlin, J. B. (1988). Network Flows. Cambridge, Mass.

Austin, P. C. (2011a). An introduction to propensity score methods for reducing the effects of confounding in observational studies. Multivariate Behavioral Research, 46(3):399–424.

Austin, P. C. (2011b). Optimal caliper widths for propensity-score matching when estimating differences in means and differences in proportions in observational studies. Pharmaceutical Statistics, 10(2):150–161.

Austin, P. C. and Stuart, E. A. (2015). Moving towards best practice when using inverse probability of treatment weighting (IPTW) using the propensity score to estimate causal treatment effects in observational studies. Statistics in Medicine, 34(28):3661–3679.

Bennett, M., Vielma, J. P., and Zubizarreta, J. R. (2020). Building representative matched samples with multi-valued treatments in large observational studies. Journal of Computational and Graphical Statistics, 29(4):744–757.

Bertsekas, D. P. (1991). Linear Network Optimization: Algorithms and Codes. MIT Press.

Bertsekas, D. P. and Tseng, P. (1988). The relax codes for linear minimum cost network flow problems. Annals of Operations Research, 13(1):125–190.

Cochran, W. G. and Rubin, D. B. (1973). Controlling bias in observational studies: A review. Sankhya: The Indian Journal of Statistics, Series A, pages 417–446.

Cole, S. R. and Stuart, E. A. (2010). Generalizing evidence from randomized clinical trials to target populations: the actg 320 trial. American Journal of Epidemiology, 172(1):107–115.

Colnet, B., Mayer, I., Chen, G., Dieng, A., Li, R., Varoquaux, G., Vert, J.-P., Josse, J., and Yang, S. (2020). Causal inference methods for combining randomized trials and observational studies: a review.

Dahabreh, I. J., Robertson, S. E., Tchetgen, E. J., Stuart, E. A., and Hernán, M. A. (2019). Generalizing causal inferences from individuals in randomized trials to all trial-eligible individuals. Biometrics, 75(2):685–694.

Degtiar, I. and Rose, S. (2021). A review of generalizability and transportability.
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., et al. (2021). Emulating randomized clinical trials with nonrandomized real-world evidence studies: first results from the RCT DUPLICATE Initiative. *Circulation*, 143(10):1002–1013.

Franklin, J. M., Pawar, A., Martin, D., Glynn, R. J., Levenson, M., Temple, R., and Schneeweiss, S. (2020). Nonrandomized real-world evidence to support regulatory decision making: process for a randomized trial replication project. *Clinical Pharmacology & Therapeutics*, 107(4):817–826.

Hansen, B. B. (2004). Full matching in an observational study of coaching for the sat. *Journal of the American Statistical Association*, 99(467):609–618.

Hansen, B. B. (2007). Optmatch: Flexible, optimal matching for observational studies. *R News*, 7(2):18–24.

Hansen, B. B. (2008). The prognostic analogue of the propensity score. *Biometrika*, 95(2):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(3):609–627.

Hernán, M. A., Alonso, A., Logan, R., Grodstein, F., Michels, K. B., Stampfer, M. J., Willett, W. C., Manson, J. E., and Robins, J. M. (2008). Observational studies analyzed like randomized experiments: an application to postmenopausal hormone therapy and coronary heart disease. *Epidemiology (Cambridge, Mass.)*, 19(6):766.

Ho, D. E., Imai, K., King, G., and Stuart, E. A. (2007). Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political Analysis*, 15(3):199–236.

Humphrey, L. L., Chan, B. K., and Sox, H. C. (2002). Postmenopausal hormone replacement therapy and the primary prevention of cardiovascular disease. *Annals of Internal Medicine*, 137(4):273–284.

Imbens, G. W. (2015). Matching methods in practice: Three examples. *Journal of Human Resources*, 50(2):373–419.

Kaplan, E. L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53(282):457–481.

Korte, B. H. and Vygen, J. (2011). *Combinatorial Optimization*, volume 1. Springer.
Lee, K., Small, D. S., and Dominici, F. (2021). Discovering heterogeneous exposure effects using randomization inference in air pollution studies. *Journal of the American Statistical Association*, pages 1–12.

Lee, K., Small, D. S., Hsu, J. Y., Silber, J. H., and Rosenbaum, P. R. (2018a). Discovering effect modification in an observational study of surgical mortality at hospitals with superior nursing. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 181(2):535–546.

Lee, K., Small, D. S., and Rosenbaum, P. R. (2018b). A powerful approach to the study of moderate effect modification in observational studies. *Biometrics*, 74(4):1161–1170.

Li, Y. P., Propert, K. J., and Rosenbaum, P. R. (2001). Balanced risk set matching. *Journal of the American Statistical Association*, 96(455):870–882.

Michels, K. B. and Manson, J. E. (2003). Postmenopausal hormone therapy. *Circulation*, 107(14):1830–1833.

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

Pimentel, S. D., Kelz, R. R., Silber, J. H., and Rosenbaum, P. R. (2015). 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(510):515–527.

Prentice, R. L., Langer, R., Stefanick, M. L., Howard, B. V., Pettinger, M., Anderson, G., Barad, D., Curb, J. D., Kotchen, J., Kuller, L., et al. (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(5):404–414.

Prentice, R. L., Langer, R. D., Stefanick, M. L., Howard, B. V., Pettinger, M., Anderson, G. L., Barad, D., Curb, J. D., Kotchen, J., Kuller, L., et al. (2006). Combined analysis of women’s health initiative observational and clinical trial data on postmenopausal hormone treatment and cardiovascular disease. *American Journal of Epidemiology*, 163(7):589–599.

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., et al. (2009). Benefits and risks of postmenopausal hormone therapy when it is initiated soon after menopause. *American Journal of Epidemiology*, 170(1):12–23.
R Core Team (2013). R: A language and environment for statistical computing.

R Core Team (2021). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rassen, J. A., Shelat, A. A., Myers, J., Glynn, R. J., Rothman, K. J., and Schneeweiss, S. (2012). One-to-many propensity score matching in cohort studies. *Pharmacoepidemiology and Drug Safety*, 21:69–80.

Rosenbaum, P. R. (1989). Optimal matching for observational studies. *Journal of the American Statistical Association*, 84(408):1024–1032.

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

Rosenbaum, P. R. (2010). *Design of Observational Studies*. Springer.

Rosenbaum, P. R. (2020). Modern algorithms for matching in observational studies. *Annual Review of Statistics and Its Application*, 7:143–176.

Rosenbaum, P. R., Ross, R. N., and Silber, J. H. (2007). Minimum distance matched sampling with fine balance in an observational study of treatment for ovarian cancer. *Journal of the American Statistical Association*, 102(477):75–83.

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

Rosenbaum, P. R. and Rubin, D. B. (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician*, 39(1):33–38.

Rossouw, J. E., Prentice, R. L., Manson, J. E., Wu, L., Barad, D., Barnabei, V. M., Ko, M., LaCroix, A. Z., Margolis, K. L., and Stefanick, M. L. (2007). Postmenopausal hormone therapy and risk of cardiovascular disease by age and years since menopause. *JAMA*, 297(13):1465–1477.

Rubin, D. B. (1979). Using multivariate matched sampling and regression adjustment to control bias in observational studies. *Journal of the American Statistical Association*, 74(366a):318–328.

Rubin, D. B. (1980). Bias reduction using mahalanobis-metric matching. *Biometrics*, pages 293–298.
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(1):20–36.

Rubin, D. B. et al. (2008). For objective causal inference, design trumps analysis. *Annals of Applied Statistics*, 2(3):808–840.

Rutter, M. (2007). *Identifying the Environmental Causes of Disease: How Should We Decide what to Believe and when to Take Action?* Academy of Medical Sciences.

Scirica, B. M., Bhatt, D. L., Braunwald, E., Steg, P. G., Davidson, J., Hirshberg, B., Ohman, P., Frederick, R., Wiviott, S. D., Hoffman, E. B., et al. (2013). Saxagliptin and cardiovascular outcomes in patients with type 2 diabetes mellitus. *New England Journal of Medicine*, 369(14):1317–1326.

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

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

Stuart, E. A., Cole, S. R., Bradshaw, C. P., and Leaf, P. J. (2011a). 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(2):369–386.

Stuart, E. A., King, G., Imai, K., and Ho, D. (2011b). Matchit: nonparametric preprocessing for parametric causal inference. *Journal of Statistical Software*.

The Women’s Health Initiative Study Group (1998). Design of the women’s health initiative clinical trial and observational study. *Controlled Clinical Trials*, 19(1):61–109.

Willett, W. C., Manson, J. E., Grodstein, F., Stampfer, M. J., and 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(11):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(3):321–333.

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(2):628–636.

Yu, R., Silber, J. H., Rosenbaum, P. R., et al. (2020). Matching methods for observational studies derived from large administrative databases. *Statistical Science*, 35(3):338–355.

Yu, R., Small, D. S., and Rosenbaum, P. R. (2021). The information in covariate imbalance in studies of hormone replacement therapy. *The Annals of Applied Statistics*, 15(4):2023 – 2042.

Zhang, B., Small, D., Lasater, K., McHugh, M., Silber, J., and Rosenbaum, P. (2021). Matching one sample according to two criteria in observational studies. *Journal of the American Statistical Association*, (just-accepted):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(500):1360–1371.