Received: 23 May 2022 | Revised: 8 February 2023 | Accepted: 20 February 2023

DOI: 10.1002/sim.9698

RESEARCH ARTICLE

Statistics in Medicine WILEY

Flexible template matching for observational study design

Ruochen Zhao | Bo Lu

Division of Biostatistics, College of Public Health, The Ohio State University, Columbus, Ohio, USA

Correspondence
Bo Lu, Division of Biostatistics, College of Public Health, The Ohio State University, 244 Cunz Hall, 1841 Neil Avenue, Columbus, OH 43210, USA.
Email: lu.232@osu.edu

Funding information
Eunice Kennedy Shriver National Institute of Child Health and Human Development, Grant/Award Number: R01HD107280; National Science Foundation, Grant/Award Number: DMS-2015552

Matching is a popular design for inferring causal effect with observational data. Unlike model-based approaches, it is a nonparametric method to group treated and control subjects with similar characteristics together, hence to re-create a randomization-like scenario. The application of matched design for real world data may be limited by: (1) the causal estimand of interest; (2) the sample size of different treatment arms. We propose a flexible design of matching, based on the idea of template matching, to overcome these challenges. It first identifies the template group which is representative of the target population, then match subjects from the original data to this template group and make inference. We provide theoretical justification on how it unbiasedly estimates the average treatment effect using matched pairs and the average treatment effect on the treated when the treatment group has a bigger sample size. We also propose using the triplet matching algorithm to improve matching quality and devise a practical strategy to select the template size. One major advantage of matched design is that it allows both randomization-based or model-based inference, with the former being more robust. For the commonly used binary outcome in medical research, we adopt a randomization inference framework of attributable effects in matched data, which allows heterogeneous effects and can incorporate sensitivity analysis for unmeasured confounding. We apply our design and analytical strategy to a trauma care evaluation study.

KEYWORDS

attributable effect, average treatment effect, poly-matching, sensitivity analysis, template matching

1 | INTRODUCTION

Observational studies provide important sources of information to infer about causal effects in clinical and health research. Compared to randomized controlled trials, the lack of randomization presents challenges in removing bias in causal effect estimation, due to observed and unobserved factors. But such design with less stringent requirement also makes it possible to address many scientific questions when randomization is not feasible or ethical. Commonly used strategies for handling measured confounding are through design adjustment, modeling, or a combination of both. The design tools include matching, stratification and weighting. The goal is to re-create a randomization-like scenario by balancing observed covariate distributions. They are usually implemented with estimated propensity scores.[1] Parametric or semi-parametric models can be used to infer causal effects, with the caveat of correct model specification. To improve

Abbreviations: ATE, overall average treatment effect; ATT, average treatment effect on the treated.

the robustness of the results, combining regression models with matching, stratification or weighting is recommended, which may have the benefit of being doubly robust.[2]

Matching is a nonparametric way of grouping subjects with similar covariates from different treatment arms together, hence to achieve balance within matched sets. With well matched data, treatment assignment may be regarded as random within each set, closely resembling a block randomization design. The advantages of matching designs are: (1) It is more robust in the sense that it balances the covariate distributions between treated and untreated groups nonparametrically, which does not rely on strong parametric assumptions for the outcome model. As an alternative to using a parametric propensity score model, researchers can consider nonparametric estimation of propensity scores, such as tree-based methods,[3] or using different distance metrics, for example, rank Mahalanobis distance, which requires little parametric assumptions.[4] (2) It resembles a randomization design, which is easily interpretable to general audience. (3) It is more objective in the sense that the causal effect inference is conducted only after good matches are established and the outcome variable never enters the matching process. (4) As a design tool, it leaves analytical options open for post-matching inference, from randomization-based nonparametric tests (more robust) to various regression analyses (more efficient). (5) Matching-based sensitivity analysis is well developed to assess the impact of hidden bias (unmeasured confounding) on causal effects based on observed data. Depending on the data structure, various types of matching design may be considered, such as pair matching, 1:k matching, variable matching, full matching or non-bipartite matching.[5]

Practically, the choice of matched design depends on the data structure. The sample size in each treatment arm and the overlap of covariate distributions may limit the design options and the causal interpretation of the matching estimator. Most matched designs lead to estimators of average treatment effect on the treated (ATT), with the exception of full matching, which is essentially a very fine way of stratification.[6] In many studies, the causal effect of interest is about the overall average treatment effect (ATE). Then most conventional matched designs can not be used. For example, if we apply a matched pairs design to a scenario with a small treated group and a large control group. Matched controls are selected in a way to mimic the distribution of the subpopulation of those who actually received the treatment. Therefore, the matched data do not provide a good answer to the causal effect of the entire population unless the treated group is a simple random sample from the population.

Another related issue is that, even if we are interested in ATT, we may still encounter a dataset with a larger treated group and a smaller untreated group. Then we cannot come up with an adequate matching estimator using the conventional design. This is because that the conventional matched design always trims off from the larger group and matches to the smaller group. A real example of such scenario was presented in Liu and her colleagues' work, where they considered borrowing historical control patients to supplement a current clinical trial of a rare disease.[7] To mimic the patient characteristics in the current trial, they tried to select the historical control patients to match the trial population. But the size of the historical control pool was smaller than that of the current trial, because these controls had to come from similar trials conducted in recent years. The authors came up with an ad hoc strategy by selecting a small representative sample from the current trial patients and matching the historical control group to the small selected sample. It seemed to work well in their study, but a general purpose methodology was not discussed.

In this article, we propose a flexible matched design to overcome the aforementioned challenges in practice with theoretical justification. It is based on the template matching strategy,[8] which was initially proposed to match across multiple groups. The basic idea is to first identify the template group with desired characteristics, then match subjects from different treatment arms of the original data set to this template group and make inference. For example, to use a matched pairs design to learn about ATE, we will first take a small and representative sample of the entire population (like a simple random sample). Using this group as the template, we will match treated and untreated subjects in the original data to form matched pairs. Because all subjects are matched to the template group, it will reflect the population described by the template, hence yield an ATE interpretation. Another contribution we make to the literature is to devise a strategy for template size selection, which determines a reasonable template size based on the average increase in total distance per unit. We also implement randomization-based inference and sensitivity analysis for template matching with binary outcomes. Under the potential outcome framework, the conventional randomization inference for non-zero effect assumes a constant additive effect for everyone. This is generally inadequate for binary data, where the outcome can only take either 0 or 1. As binary outcomes are very common in medical research, Rosenbaum proposed an attributable effect framework for conducting randomization inference with binary data.[9] It captures the effect attributable to the treatment in the sense that it would not have been observed had the individual not been exposed to treatment. It allows for heterogeneous effects in the target population, where part of the population may experience the effect and others do not. The sensitivity analysis for unmeasured confounding can also be incorporated easily into the randomization inference procedures. We implement this estimation strategy for our matched data, examining mortality outcomes in trauma care evaluation.

For the rest of this article, we will illustrate how to use our flexible matched design: (1) to construct a pair design to estimate ATE with improved matching quality via the triplet matching algorithm; (2) to form matched pairs to estimate ATT when the treated group has a larger size. We will also discuss how to determine the sample size of template group to best represent the target population. Specifically, in Section 2, we provide theoretical justification on how the proposed template matching can unbiasedly estimate ATE and ATT. We also discuss strategies for template size selection in practice. In Section 3, we conduct simulation studies to justify the empirical performance of our proposed design. A real data example of trauma care evaluation is presented in Section 4, where we briefly review and implement the attributable effect estimation strategy. We also conduct a sensitivity analysis to assess the impact of hidden bias, and examine its robustness by varying the template size. We conclude with summary and discussion in Section 5.

2 | TEMPLATE MATCHING DESIGN

2.1 | Overview of template matching design

The idea of template matching was first proposed by Silber and his colleagues, where they tried to come up with a fair audit of quality and cost for Medicare patients across 217 hospitals.[10] The direct comparison of all patients in these hospitals is problematic, as patient characteristics are very different in different hospitals. Silber and colleagues, instead, chose a small template sample of 300 patients with characteristics representative to the patient population of interest for the audit. Within each hospital, 300 patients were individually paired to the 300 patients in the template. With successful matches, all hospitals were evaluated fairly on patients with similar backgrounds. Bennett et al extended this idea to representative matching with multi-valued treatments.[8] Unlike matching with two treatment arms, matching with three or more treatments is computationally intractable and the optimal algorithm is not available. By choosing a common template and matching every treatment arm to this template, Bennett and colleagues converted a multi-arm matching problem to a simpler repeated two-arm matching problem.

Generally, template matching refers to a class of designs that uses a small but representative template as the benchmark and matches all other groups to this template. After matching, the inference and interpretation are based on the population characteristics of such template sample. It adds a lot of flexibilities in matching practice and avoids the restriction in the conventional matching that all subjects in a particular group must be used. It is particularly useful when matching with a large number of treatment groups or population characteristics are different among different groups with no single group representing the population of interest, as discussed above. In fact, many conventional matched designs can be considered as special cases of template matching, for example, a design selecting control subjects to match to the treated group is literally using the whole treated group as a template. In this article, we want to further extend the template matching to address some commonly encountered issues in two-arm matching, such as estimating ATE or estimating ATT when the treatment group has a bigger sample size than the control group. Typically, template matching includes two steps:

Step 1: A template sample is selected depending on the goal of matching or the purpose of the study. The template sample should be representative to the population of interest and have a smaller sample size than that of any original treatment group to facilitate matching.

Step 2: Matching without replacement will be conducted between the template sample and the original treatment groups to create a randomization-like scenario. Optimal bipartite matching algorithm may be used if the matching is between two groups and some suboptimal matching algorithm may be considered if the matching involves more than two groups.

Technically, any distance metric can be used in template matching, such as propensity score distance, Mahalanobis distance and so forth. Matching quality needs to be carefully examined before moving forward to inference. The next two subsections provide more details about the template matching design for the two practical issues that we would like to address in this article. To fix the terminology, we use treatment and control to denote the two possible treatment arms for the rest of the article. They are used in the broad sense as treatment means an active/experimental treatment or exposure to the causal agent, and control means no treatment, standard care or no exposure.

2.2 Template matching for ATE estimation

For notation simplicity, we discuss template matching to estimate ATE in a binary treatment setting, but the idea of template matching can be easily extended to multi-valued treatment settings. Suppose there are n_T treated subjects, $\mathcal{T} = \{\tau_1, \dots, \tau_{n_T}\}$ and n_C controls $C = \{\gamma_1, \dots, \gamma_{n_C}\}$. The whole population P is the combination of treatment group $\mathcal{T}$ and control group C, $P = \{\mathcal{T} \cup C\} = \{\tau_1, \dots, \tau_{n_T}, \gamma_1, \dots, \gamma_{n_C}\} = \{p_1, \dots, p_{n_T}, \dots, p_{n_T+n_C}\}$. Each subject has a vector of observed covariates, denoted as $\mathcal{X}$, which is used in calculating the distance metric for matching. The population of interest is the whole population P when the target estimand is ATE, so that a representative template sample should be first selected from the whole population P. Suppose a template with sample size m is drawn ($m < \min(n_T, n_C)$), and we denote it as $S = \{p_{S1}, \dots, p_{Sm}\}$. Then we can carry out the template matching as shown in Figure 1. Specifically, the template group is first identified from the whole population, then treatment and control groups are matched to mimic the template distribution via triplet matching. Denoting $M = \{(\tau_{S1}, \gamma_{S1}), \dots, (\tau_{Sm}, \gamma_{Sm})\}$ as the final matched set of treated and control unit pairs, we can show that the template matching produces unbiased ATE estimation, which follows the similar idea of Rosenbaum and Rubin's seminal paper.[1] We need two general assumptions for causal inference, namely, stable unit treatment value assumption (SUTVA) and strongly ignorable treatment assignment assumption (SITAA).[11] The SUTVA assures no interference and the treatment value is consistent. The SITAA assures unconfoundedness given covariates $\mathcal{X}$ and good overlap between treated and control groups. We also assume that the treated and control groups are independent draws from their respective population distributions. Suppose each subject has two potential outcomes $\{Y^1, Y^0\}$, which are the responses under treatment ($T = 1$) or control ($T = 0$) respectively. With SITAA given covariates $\mathcal{X}$, we also have the strong ignorability given any balancing score $b(\mathcal{X})$, which is based on the study population of the original treated and control groups. The following result establishes the unbiased ATE estimation with the exact template matching. According to Rosenbaum and Rubin's framework, we match on true population balancing scores, or it is applicable to large samples when the balancing scores can be consistently estimated. The exact template matching creates a matched triplet with one unit from the template group, one from the treated group and one from the control group, and all of them have the same value of the balancing score. "Exact" means that three units in each triplet share the same value of the balancing score, and "template matching" means that, as a group, the matched treated units have the same balancing score distribution as the template group, and so do matched control units. Then, we can ignore the template group and focus on the matched pairs of treated and control units to obtain an unbiased estimate of ATE.

Result 1. *Assume SUTVA, SITAA, and the template group is representative of the entire population, then, following the exact template matching, the mean difference between treated and control groups in the matched sample is an unbiased estimator for the ATE, that is,*

$$E_P[Y^1 - Y^0] = E_M[Y|T = 1] - E_M[Y|T = 0]. \tag{1}$$

Proof. Since the template sample, denoted by S, is representative of the entire population, denoted by P, it implies:

$$E_P[Y^1 - Y^0] = E_S[Y^1 - Y^0].$$

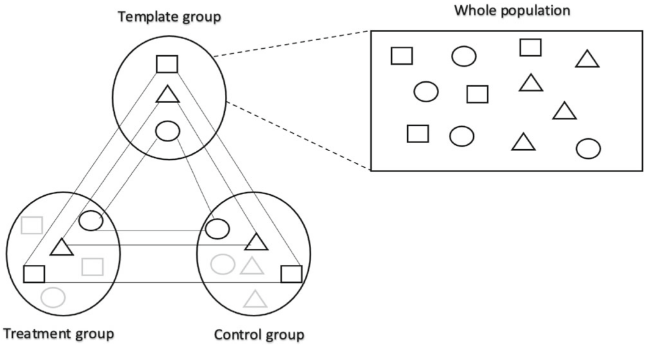

FIGURE 1 Template matching for ATE estimation. The dashed lines indicate the draw of template group and the solid lines indicate the triplet matching among three groups

Matching on $b(\mathcal{X})$ across all three groups ensures the template sample, the matched treated group and the matched control group have the same distribution of $b(\mathcal{X})$ as the target population. For any matched pair with a given value of the balancing score, under the strongly ignorable assumption, we have

$$\begin{aligned} E[Y^1 - Y^0|b(\mathcal{X})] = E[Y^1|b(\mathcal{X})) - E(Y^0|b(\mathcal{X})] = E[Y^1|T = 1, b(\mathcal{X})] \\ - E[Y^0|T = 0, b(\mathcal{X})] = E[Y|T = 1, b(\mathcal{X})] - E[Y|T = 0, b(\mathcal{X})]. \end{aligned}$$

It implies that the expected difference in response between the two units in the matched pair equals the ATE at the particular balancing score value. Then when we take the expectation over the entire distribution of $b(\mathcal{X})$ in the matched sample, we have

$$E_M[Y^1 - Y^0] = E_M[E(Y^1 - Y^0|b(\mathcal{X}))] = E_M\{E[Y|T = 1, b(\mathcal{X})] - E[Y|T = 0, b(\mathcal{X})]\} = E_M[Y|T = 1] - E_M[Y|T = 0].$$

Since the template matching ensures that the template group (S) and the matched treated and control groups (M) have the same balancing score distribution, we get

$$E_P[Y^1 - Y^0] = E_S[Y^1 - Y^0] = E_M[Y^1 - Y^0] = E_M[Y|T = 1] - E_M[Y|T = 0].$$

There are two ways of implementing template matching with three groups. The first way follows Bennett's idea by conducting bipartite matching twice, once between the original treatment group and the template sample, and once between the original control group and the template sample.[8] The treatment and control groups are indirectly linked through the template group. The major advantage of this implementation is simplicity. However, since it does not control the distance between treatment and control groups directly, it may produce matched treated-control pairs with total distance larger than the optimal one. To improve upon this, we modify the matching algorithm by including the distance between treatment and control groups explicitly.

The improved way of implementation is via triplet matching. Triplet matching algorithm was proposed by Nattino and colleagues to form better matched triplets among three groups by considering the three-way distance.[12] The three-way distance within a matched triplet is defined as $d^3(\tau_{T_k}, \gamma_{C_k}, p_{S_k}) = d^2(\tau_{T_k}, \gamma_{C_k}) + d^2(\tau_{T_k}, p_{S_k}) + d^2(\gamma_{C_k}, p_{S_k})$, where $d^2(.,.)$ can be any two-way distance and τ_{T_k}, γ_{C_k}, p_{S_k} are subjects from treatment, control and template groups, respectively. In our study, we consider the Euclidean distance, so that the three-way distance $d^3(\tau_{T_k}, \gamma_{C_k}, p_{S_k})$ represents the perimeter of the triangle defined by the three subjects from a matched triplet, as illustrated in Figure 1. The key difference between our approach and Bennett's approach above is that the latter only considers two side lengths of the triangle, not all three. With the distance between the treatment and control groups being included directly, we expect triplet matching to improve the similarity between those two groups. The sum of the within-triplet distances is $D(\mathcal{T}_r) = \sum_{k=1}^{m} d^3(\tau_{T_k}, \gamma_{C_k}, p_{S_k})$. The optimal triplet matching that minimizes $D(\mathcal{T}_r)$ is not available in polynomial time. Nattino and colleagues showed that their polynomial-time algorithm is near-optimal in the sense that the final matched triplets have a total distance no more than twice of the optimal result.[12] To implement this algorithm, we use the matched triplets from the first method discussed above as a starting point. Then we fix the selected template sample S, and search in the original treatment group $\mathcal{T}$ and control group C to see if we can find a new matched triplet $\mathcal{T}_r' = \{(\tau_{T_k}', \gamma_{C_k}', p_{S_k}')\}$ with a smaller total distance. A new set of matched triplets with a substantially smaller distance may lead to improved balance between the treatment and control groups. When there are more than two treatment groups, the general poly-matching algorithm can be used to create matched sets. Poly-matching refers to the matching design that generates matched sets with at least one subject from each treatment group.[13] Triplet matching is a special case of poly-matching. If the matching process achieve desirable balance, both matched treated and control subjects mimic the distribution of the template group, hence the distribution of the whole population P. Then we can safely ignore the template group and continue the statistical analysis with the matched subjects from the treatment and control groups, which should reflect the causal effect of ATE.

The first step of template matching is to select a representative template sample. Though a simple random draw from the population P to select the template may work in theory, practically, it is desirable to draw multiple samples and pick the most representative one to reduce the chance error that may distort the distribution.[8] Specifically, we draw n simple random samples of size m from the whole population

P, denoted as $\mathcal{S} = \{S_1, S_2, \dots, S_n\}$, $S_i = \{p_{i1}, p_{i2}, \dots, p_{im}\}$, $i = 1, \dots, n$. For each random sample S_i, we calculate the rank Mahalanobis distance for each pair of data points between S_i and the population P, resulting in a $m \times (n_T + n_C)$ distance matrix W_i. The sum of all elements in the matrix W_i is computed, denoted as w_i (a measure of total distance). The sample S_t with the smallest total distance is regarded as being closest to the whole population. This sample will be used as the template to proceed with the matching process. ■

2.3 | Template matching for ATT estimation with a larger treatment group

When the target causal estimand is ATT, the conventional bipartite matching design requires that the sample size in the treatment group is less than that of the control group. This is because the matching algorithm always selects subjects from the larger group to match to the smaller group. Practically, however, there is no guarantee that the treatment group is always smaller. Therefore, the implementation of conventional matching design might be problematic. Liu and colleagues presented a real-world example of matching with a larger treatment group in a clinical trial using historical control borrowing to boost the sample size.[7] The clinical trial was designed to evaluate the efficacy of a new treatment for a very rare tumor. Given the rareness of the disease, it was very difficult to enroll a large number of patients. Like many clinical studies for rare cancer, historical control patients from previous trials were borrowed to facilitate the study. To ensure the comparability of historical subjects, a small pool of control patients were identified based on similar inclusion and exclusion criteria, prior treatment, cancer types and endpoint availability. The historical controls were further selected to match the distribution of current trial patient population, which created an issue of matching a smaller group to a larger group. Liu and colleagues drew a smaller random sample from the current trial population and matched the historical control group to it. We will formalize their idea in terms of template matching in this subsection.

Suppose that the population of interest is treatment group $\mathcal{T} = \{\tau_1, \dots, \tau_{n_T}\}$ and the sample size of the control group is small, that is, $n_C < n_T$, then a representative template sample should be first selected from $\mathcal{T}$. Suppose a template with sample size m is drawn ($m < n_C$), denoted as $S = \{p_{S1}, \dots, p_{Sm}\}$. Since S has a smaller sample size than the control group, conventional bipartite matching can be conducted from the control group to the template group, as shown in Figure 2 (the dashed lines indicate the first-step selection of the template sample from the treatment group). The exact template matching creates a matched pair with one unit from the template group and one unit from the control group, and both have the same value of the balancing score. “Exact” means that two units in each pair share the same value of

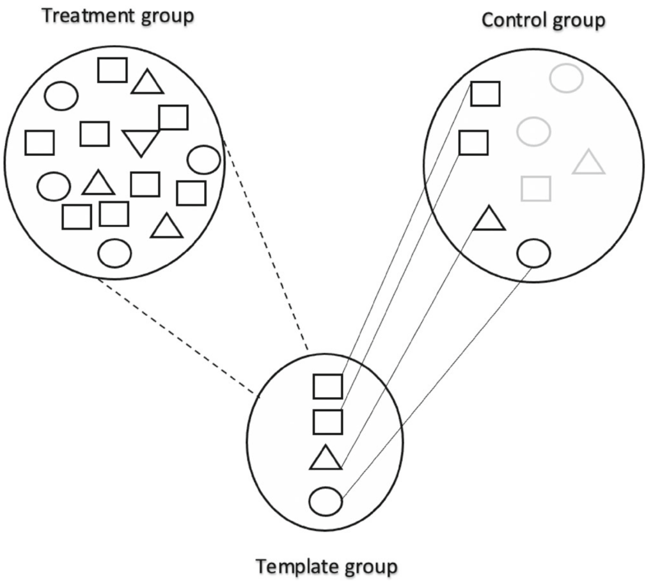

FIGURE 2 Template matching for ATT estimation. The dashed lines indicate the draw of template group and the solid lines indicate the two-group matching between template and control groups

the balancing score, and "template matching" means that, as a group, the matched control units have the same balancing score distribution as the template group. Matched pairs are denoted as $M = \{\gamma_{C_k}, p_{S_k}\}, 1 \leq k \leq m$. We have m pairs in total since all subjects in the template group are used. If the template group is representative of the original treatment group and matched pairs have desirable balance, both the template group and the matched control group mimic the distribution of treatment group $\mathcal{T}$. Then inference based on the matched data will reflect the causal effect of ATT, as shown in the following result.

Result 2. *Assume SUTVA, SITAA, and template sample is representative of the treated population, then, following the exact template matching, the mean difference between treated and control groups in the matched sample is an unbiased estimator for the ATT, that is,*

$$E_T[Y^1 - Y^0] = E_M[Y|T = 1] - E_M[Y|T = 0]. \tag{2}$$

Proof. Since the template sample, denoted by S, is representative of the treated population, denoted by T, it implies:

$$E_T[Y^1 - Y^0] = E_S[Y^1 - Y^0].$$

Matching on $b(\mathcal{X})$ ensures the template sample and the matched control group have the same distribution of $b(\mathcal{X})$. For any matched pair with a given value of the balancing score, under the strongly ignorable assumption, we have

$$\begin{aligned} E[Y^1 - Y^0|b(\mathcal{X})] = E[Y^1|b(\mathcal{X})] - E[Y^0|b(\mathcal{X})] = E[Y^1|T = 1, b(\mathcal{X})] \\ - E[Y^0|T = 0, b(\mathcal{X})] = E[Y|T = 1, b(\mathcal{X})] - E[Y|T = 0, b(\mathcal{X})]. \end{aligned}$$

It implies that the expected difference in response between the two units in the matched pair equals the ATE at the particular balancing score value. Then when we take the expectation over the entire distribution of $b(\mathcal{X})$ in the matched sample, we have

$$E_M[Y^1 - Y^0] = E_M[E(Y^1 - Y^0|b(\mathcal{X}))] = E_M\{E[Y|T = 1, b(\mathcal{X})] - E[Y|T = 0, b(\mathcal{X})]\} = E_M[Y|T = 1] - E_M[Y|T = 0].$$

Since the template matching ensures that the template group (S) and the matched treated and control groups (M) have the same balancing score distribution, we get

$$E_T[Y^1 - Y^0] = E_S[Y^1 - Y^0] = E_M[Y^1 - Y^0] = E_M[Y|T = 1] - E_M[Y|T = 0].$$

To select a representative template sample, we use a similar approach as described in the previous subsection. We draw n simple random samples of size m from the treatment group $\mathcal{T}$, denoted as $\mathcal{S} = \{S_1, S_2, \ldots, S_n\}$, $S_i = \{p_{i1}, p_{i2}, \ldots, p_{im}\}$, $i = 1, \ldots, n$. The rank Mahalanobis distance is calculated between each S_i and the treatment group $\mathcal{T}$, resulting in a $m \times n_T$ distance matrix W_i. The sum of all elements in the matrix W_i is computed, denoted as w_i (a measure of total distance). The sample S_t with the smallest total distance is regarded as being closest to the treatment group. ■

2.4 | Strategies for template size selection

In Sections 2.2 and 2.3, we describe the implementation of template matching with a fixed template sample size of m. In practice, we may face the question on how to determine a reasonable template size. There are two ways to approach it. The first approach is design-driven, where the m is selected as part of the design parameters. Liu and colleagues pre-specified $m = 30$ for the historical control cohort given that their current randomized trial had 60 treated and 30 control patients.[7] Therefore they would have equal-sized treated and control groups for further analysis. The second approach is data-driven, where is more suitable for large observational studies. Similar to the issue of matching with a subset of sample discussed in Rosenbaum's work,[14] there are two competing goals in determining the adequate template sample size m. On one hand, we want to make m as large as possible to increase the efficiency of statistical inference. On the other hand, we want closely matched treatment and control groups. A very large m may lead to matched sample with poor balance, especially when covariate distribution of the two groups have moderate but not great overlap. Specifically, in estimating ATE, the first goal demands a larger m. While the second goal is to keep the total distance, $w(S) = \sum_{k=1}^{m} \delta(\tau_{T_k}, \gamma_{C_k})$, small to ensure well-balanced matched pairs. Numerically, the two goals may pull in opposite directions, since increasing template size also increases the total distance between matched pairs, which means when $m' > m$, we have $w(S') > w(S)$. Thus a reasonable template size selection strategy needs to achieve a balance between the two goals. Rosenbaum considered a selection criterion on the average increase in total distance between two sample size scenarios m' and m.[14] Assuming $m' > m$, with a given threshold $\tilde{\delta}$, we prefer m to m', if

$$\frac{w(S') - w(S)}{m' - m} > \tilde{\delta}. \tag{3}$$

It implies, moving from m to m', template group size increases by $m' - m$ and the total distance between matched pairs increases by $w(S') - w(S)$, then we would prefer the smaller template size m provided that the average increase in total distance per unit, $\frac{w(S')-w(S)}{m'-m}$, is larger than the ideal threshold. In other words, it is preferable not to sacrifice the loss in similarity between the matched groups in exchange for a larger number of matched pairs. By adopting a similar strategy, we first identify the lower bound, l, and the upper bound, h, for the possible template size, for example, $h \leq \min(T, C)$. Then, we select q potential options for template sample size $\{m_1, m_2, \dots, m_q\}$ in the range of $[l, h]$ in ascending order, find the most representative template samples $\{S_1, S_2, \dots, S_q\}$ and conduct template matching for each sample size option. The optimal template size is chosen as $m = \max\{m_i : \frac{w(S_i)-w(S_{i-1})}{m_i - m_{i-1}} \leq \tilde{\delta}, 2 \leq i \leq q\}$. This is the biggest size we can have without letting the average increase in total distance exceed a desired value. As Rosenbaum suggested, $\tilde{\delta}$ may be chosen as a certain quantile of the pairwise distance distribution from the target population.[14] As shown later in the real data analysis, we use the 25th percentile as the threshold to select the template size for the trauma center patients.

3 | SIMULATION STUDIES

We conduct a series of simulation studies to examine the performance of template matching for estimating the causal effect with both continuous and binary outcomes, with a primary focus on bias, as matching is designed to remove bias due to observed confounding. Section 3.1 focuses on ATE and Section 3.2 focuses on ATT when the treated group has a larger sample size.

3.1 | ATE estimation

We consider the template matching strategy discussed in Section 2.2 to estimate the ATE. For comparison, we also run full matching, which is known to provide ATE estimation. Full matching literally stratifies the data to the finest possible groups, ensuring at least one treated and one control subject in each stratum.[6] For all simulation scenarios, we generate 2000 datasets and the sample size for each dataset is 1000. The following two subsections cover continuous outcome and binary outcome respectively.

3.1.1 | Continuous outcome

Data-generating process

Three covariates are considered in the study. X_1 and X_3 are generated from the normal distribution and X_2 is generated from the Bernoulli distribution.

$$X_1 \sim N(10, 5), X_2 \sim Bern(0.4), X_3 \sim N(0, 1).$$

The treatment assignment is generated from a logistic regression model, and the treatment prevalence is 20%.

$$\text{logit}(P(T = 1)) = -0.7 + \log(0.9)X_1 + \log(1.5)X_2 + \log(2)X_3. \tag{4}$$

The outcome model is $Y \sim N(\mu, 3^2)$, and we have three settings for the outcome model, where true ATE are 5.5, 1.7, and 0 respectively, for large, small, and no effect scenarios. The outcome model is designed to include heterogeneous effects modified by covariates to make ATE and ATT different.

$$\begin{aligned}\text{Large effect:}\quad & \mu = 2 + (3 + 3 * I(X_1 > 10) + 2.5 * X_2) * T + \log(1.2)X_3,\\ \text{Small effect:}\quad & \mu = 2 + (1 + 1 * I(X_1 > 10) + 0.5 * X_2) * T + \log(1.2)X_3,\\ \text{No effect:}\quad & \mu = 2 + (1 + 1 * I(X_1 > 10) + 0.5 * X_2) * 1 + \log(1.2)X_3.\end{aligned} \tag{5}$$

Statistical analyses in simulated datasets

For ATE estimation, the target population includes everyone, both treated and control subjects, so the template is chosen to mimic the entire study population. We first estimate the propensity score using all 1000 subjects, via a logistic regression model to regress treatment assignment on the three covariates X_1, X_2, and X_3. Optimal full matching on the estimated propensity score is conducted in R using package "optmatch" and command "estimand = ATE," which includes all subjects in the matched sample by design. For template matching, template sizes are chosen as 60 or 120 to explore the impact of small and large sizes. For a template sample of size n, 25 random samples are drawn from each dataset. From these 25 random samples, we select the sample that is closest to the corresponding dataset as the template group based on the rank Mahalanobis distance. Treatment and control groups are then matched separately to the template sample, using the optimal bipartite matching algorithm based on the estimated propensity scores. The last step is to further improve covariate balance among the three groups (template, treated, and control) via triplet matching. This is important as the triplet matching controls directly for the discrepancy between treatment and control groups, while the bipartite matching does not. Optimal bipartite matching is implemented using "optmatch" package in R and triplet matching is implemented using "polymatching" package in R.

Table 1 presents the results for continuous outcome analysis. We apply a simple linear regression model to estimate treatment effect in matched data. Let θ denote the true treatment effect of ATE, and $\hat{\theta}_i$ denote the estimated ATE in each dataset. The percentage of mean relative bias is estimated as $\frac{\frac{1}{2000}\sum_{i=1}^{2000}(\hat{\theta}_i - \theta)}{\theta} \times 100$ and reported in the column "%Bias," and for the null effect scenario, we calculate overall bias instead, which is $\frac{1}{2000}\sum_{i=1}^{2000}(\hat{\theta}_i - \theta)$. "SD(ROB)" column shows the robust SE estimated in the linear regression model by accounting for the potential correlation introduced by matching. "%CI" column reports the empirical coverage of the nominal 95% confidence interval (CI) across 2000 datasets. "SD(EMP)" column shows the empirical SE for comparison purpose. Overall, template matching and full matching have similar performance in terms of bias and the empirical coverage of 95% confidence interval. Template matching with size of 60 has smaller bias but larger SE than size of 120. This is likely due to the fact that it is easier to find good matches for a smaller template sample, but the small sample size increases the variance. The variance estimates are valid as the model based variance is very close to the empirical variance.

TABLE 1 Continuous outcome with large, small, and null effect

	True ATE	%Bias	SD(ROB)	%CI	SD(EMP)
Temp (60)	5.5	−0.052	1.670	95.7	1.639
Temp (120)	5.5	−0.658	1.178	95.0	1.177
Full	5.5	0.464	1.056	94.9	1.053
Temp (60)	1.7	0.904	1.651	94.8	1.612
Temp (120)	1.7	0.963	1.408	95.0	1.405
Full	1.7	0.417	1.039	94.1	1.054
Temp (60)	0	0.018[a]	1.652	94.6	1.639
Temp (120)	0	0.000[a]	1.166	94.6	1.181
Full	0	−0.009[a]	1.038	94.7	1.070

[a] %Bias for null effect scenario is the overall bias.

3.1.2 | Binary outcome

Data-generating process

For binary outcomes, we also consider three covariates, generated from either normal distribution or Bernoulli distribution:

$$X_1 \sim N(2,1), X_2 \sim Bern(0.4), X_3 \sim N(0,1).$$

The treatment assignment model is shown below, and the treatment prevalence is set to 20%.

$$\text{logit}(P(T=1)) = -1.4 + \log(0.9)X_1 + \log(1.5)X_2 + \log(2)X_3. \tag{6}$$

We use risk difference (RD) as the outcome measure to avoid the noncollapsibility issue of odds ratio. We consider three treatment effect scenarios for the outcome model, where true RD is large, small or zero. The outcome model with large and small RD's are set up as:

$$\text{logit}(P(Y=1)) = -3.532 + 0.3 * X_1 + X_2 + X_3 + (\tau_0 + 0.3 * I(X_1 > 2) + 0.25 * X_2) * T, \tag{7}$$

where τ_0 is a parameter to determine the true treatment effect. τ_0 is set to 1.256 for large effect (RD = 0.2) and set to 0.594 for small effect (RD = 0.1) respectively, using the Monte Carlo integration approach by Austin.[15]

The outcome model for the null effect case is simple as below:

$$\text{logit}(P(Y=1)) = -3.532 + 0.3 * X_1 + X_2 + X_3. \tag{8}$$

Statistical analyses in simulated datasets

We implement the same template matching and full matching algorithm as described in continuous outcome subsection since matching process does not involve outcome variables.

For the ith round of simulation, we fit a simple logistic regression model to the matched data to estimate treatment effect. For example, with template size of 60, there are 120 matched subjects in total.

$$\text{logit}[P(Y_{ij}=1)] = \alpha_i + \beta_i * T_{ij}, i = 1, \dots, 2000, j = 1, \dots, 120. \tag{9}$$

For each subject, we compute the predictive probabilities of the outcome under treatment and control conditions, denoted as $\hat{p}_{ij,1}$ and $\hat{p}_{ij,0}$ respectively. Therefore, the estimated RD in the ith simulated dataset is $\hat{\theta}_i = \sum_{j=1}^{120}(\hat{p}_{ij,1} - \hat{p}_{ij,0})$. "%Bias" is calculated using the same formula as the continuous outcome analysis. "SD(DEL)" is the SE estimate using the delta method, which calculates the variance of $\hat{\theta}$ as a function of the estimated parameters α and β. "%CI" is the empirical coverage of the 95% confidence interval (CI), based on SD(DEL), across 2000 datasets. "SD(EMP)" is the SE of the 2000 point estimates of $\hat{\theta}_i$, which measures the empirical variability of the estimator. Full matching results are analyzed using a similar approach, the only difference is we use a weighted logistic regression model since each matched set may have different sizes. Table 2 shows that template matching with template size 60 has better performance than template size 120 regarding relative bias and confidence interval coverage. Template matching has similar relative bias as full matching when the template size is 60, but one disadvantage of full matching is that it tends to underestimate the SE of risk differences, which leads to under coverage of the 95% CI.

3.2 | ATT estimation with a larger treated group

We design a simulation analysis to evaluate the performance of template matching in estimating ATT with a larger treatment group. We consider two sets of simulations: one for continuous outcome and the other for binary outcome. In each of the considered scenarios, we generate 2000 datasets and the sample size for each dataset is 1000.

TABLE 2 Binary outcome with large, small, and null effect

	True RD	%Bias	SD(DEL)	%CI	SD(EMP)
Temp (60)	0.2	−0.691	0.071	96.7	0.066
Temp (120)	0.2	4.337	0.051	96.5	0.048
Full	0.2	0.891	0.034	90.8	0.039
Temp (60)	0.1	−2.308	0.064	95.7	0.060
Temp (120)	0.1	4.646	0.047	95.4	0.045
Full	0.1	−0.065	0.030	91.2	0.033
Temp (60)	0	0.000[a]	0.055	95.5	0.053
Temp (120)	0	0.003[a]	0.040	96.0	0.038
Full	0	0.001[a]	0.024	93.7	0.026

[a] %Bias for null effect scenario is the overall bias.

3.2.1 | Continuous outcome

We consider three covariates X_1, X_2, and X_3, which have the same distribution as the ATE simulation for the continuous outcome case in Section 3.1.1. The treatment assignment model is shown below, and treatment prevalence is set to about 70% implying more treated subjects than controls.

$$\text{logit}(P(T = 1)) = 5.68 + \log(0.65)X_1 + \log(1.5)X_2 + \log(3)X_3. \tag{10}$$

The outcome model is $Y \sim N(\mu, 1.5^2)$, and the empirical ATT estimation in treated group is 3.035415.

$$\mu = 2 + (2 + 3 * I(X_1 > 10)) * T + 2.5 * X_2 + \log(1.2)X_3. \tag{11}$$

To determine the empirical truth of ATT, we first generate 5000 datasets with sample size 1000, and the jth subject in the ith dataset has treatment status T_{ij} and outcome Y_{ij}. Then, for each subject we generate T'_{ij} with a flipped treatment status of T_{ij} (ie, $T'_{ij} = 1$ if $T_{ij} = 0$, and $T'_{ij} = 0$ if $T_{ij} = 1$), and determine the outcome Y'_{ij} by the same model except changing T_{ij} to T'_{ij}. The empirical true ATT is calculated as $\frac{1}{5000}\sum_{i=1}^{5000}\frac{1}{1000}\sum_{j=1}^{1000}\mathbf{1}(T_{ij} = 1)(Y_{ij} - Y'_{ij})$, where $\mathbf{1}(T_{ij} = 1) = 1$ if $T_{ij} = 1$, otherwise $\mathbf{1}(T_{ij} = 1) = 0$. The empirical truth of ATT is 3.035 in our simulation.

In this part of the simulation study, we examine the performance of template matching under a series of template size scenarios $\{30, 40, 60, 80, 100, 120\}$. This is to get more insights into the impact of template size on bias. Conceivably, the template size cannot be too small or too big. When it is too small, the bias tends to be large as the template sample does not capture the full characteristics of the treated population. When it is too big, the matching quality tends to be comprised, in turn, it increases the bias. Matching quality is assessed based on the relative bias of estimating ATT and the empirical 95% confidence interval coverage.

Table 3 presents the results of ATT estimation for continuous outcome analysis. We implement a similar procedure to estimate ATT and the robust SD, and calculate the relative bias as described in Section 3.1.1. The "Bias%" column shows that the relative biases are different as the template size changes. It drops below 1% when the template size is 80 or more. The empirical coverage of confidence intervals is close to 95%, but small template size scenarios tend to be a bit further away from the nominal value. All these imply that a properly selected template size may improve the performance of template matching.

3.2.2 | Binary outcome

We use the same three-covariate setup for binary outcomes as described in Section 3.1.2. The treatment assignment model is shown below, and treatment prevalence is about 70%.

$$\text{logit}(P(T = 1)) = 0.8 + \log(0.9)X_1 + \log(2.5)X_2 + \log(2)X_3. \tag{12}$$

TABLE 3 ATT estimation in continuous outcome

ATT = 3.035	Bias%	SD (ROB)	CI%
Temp (30)	−5.095	0.722	93.9
Temp (40)	−3.978	0.624	93.7
Temp (60)	−2.039	0.508	95.9
Temp (80)	−0.858	0.439	95.5
Temp (100)	0.147	0.393	96.0
Temp (120)	0.343	0.359	95.6

TABLE 4 Risk difference estimation in binary outcome

RD = 0.230	Bias%	SD (DEL)	CI%
Temp (30)	−1.245	0.116	96.1
Temp (40)	−0.774	0.101	95.7
Temp (60)	−1.950	0.083	96.3
Temp (80)	2.054	0.072	96.3
Temp (100)	3.587	0.064	95.2
Temp (120)	6.086	0.058	94.8

The outcome model can be expresses as below, where the empirical truth of risk difference (RD) in treated group is 0.230. The empirical RD is calculated in the same way as the continuous outcome scenario.

$$\text{logit}(P(Y = 1)) = -2.717 + 0.3 * X_1 + X_2 + X_3 + (-0.150 + 3 * X_2) * T. \tag{13}$$

We repeat the simulations for a series template size scenarios of $\{30, 40, 60, 80,100, 120\}$ with 2000 datasets. Relative bias of estimating true risk difference and empirical 95% confidence interval coverage are reported as below.

Table 4 provides the results of risk difference estimation in binary outcome scenario. A simple logistic model is used to estimate the risk difference and the SE is estimated using the delta method as described in Section 3.1.2. Again, the results suggest that the performance of template matching vary as the template size changes. The relative bias becomes large for bigger template sizes. The size of 40 seems to offer a good balance between bias and variance.

4 | REAL DATA EXAMPLE WITH TRAUMA CARE EVALUATION

4.1 | Causal effect on trauma patient mortality

Traumatic injury is the leading cause of death for people under the age of 45 in the U.S. and worldwide. Approximately 5.8 million people die each year because of injuries.[16] In U.S., Patients with traumatic injuries can be either admitted to trauma centers (TC) or non-trauma centers (NTC) for care. Trauma centers are supposed to offer better care with a larger patient volume and higher level of expertise, while non-trauma center have limited resources and often lack specialty doctors. This makes the more severely injured patients are more likely to be sent to trauma centers. But the actual decision involves multiple practical factors, with the distance to the nearest TC or NTC being an important one. Therefore, it is quite challenging to evaluate the performance of TC and NTC in the real world setup.[17] The Nationwide Emergency Department Sample (NEDS) is a nationally representative dataset designed by the Agency for Healthcare Research and Quality to enable analyses of emergency department (ED) utilization patterns (this data set is available subject to third party restriction). The National Trauma Data Standard Patient Inclusion Criteria is used to define trauma. Since children and older adults may respond to the treatment differently, we limit our analysis to adults aged 18-64. In this analysis, we use five years of data from the NEDS.

The primary outcome of interest is emergency department mortality. The injury severity score (ISS) is one of the most important characteristics for trauma patients, as it correlates with mortality, morbidity, and other clinical outcomes. The score ranges from 1 to 75, with 75 being the most severe. Because the majority of the emergency department deaths are in the group of patients with higher scores (injury severity score >= 25), we focus our analyses on trauma patients in this group. Other covariates include sex, age, comorbidity of chronic conditions, multiple injuries, median household income by zip code, expected primary payer, and urban-rural designation for patient's county of residence. More detailed description of this dataset is provided in the work by Vickers et al,[18] Shi et al,[19] and Nattino et al.[12] Out of the 21 855 patients included in the dataset, 16 541 (75.7%) patients were admitted to TC, and 5314 (24.3%) patients were admitted to NTC.

Since TCs are much more resource-craving than NTCs, a natural policy relevant question is whether TC has the advantage of saving patient lives to justify the high cost. So the causal question is what would have happened to those patients treated at TC had they been treated at NTC. In other words, for TC patients, we would like to know their potential outcomes had they been admitted to NTC. This calls for an ATT estimation as we are primarily interested in evaluation of TC patient population. Because there are more TC patients than NTC patients, the conventional two group matching cannot be applied and we implement the proposed template matching to select NTC patients to mimic a TC population. To select the template size, we first identify a range of sample sizes from 1500 to 2600, which corresponds to 30% to 50% of the NTC patient population to ensure that we have enough controls to choose from. With an increment of 100 patients, we calculate the total rank Mahalanobis distance and the average distance increase (from the template with the next smaller size) for each template size, as described in Section 2.4. The detailed information is summarized in Table 3 in the online supplementary material. To pick the ideal template size, we try to balance between a bigger sample size (for efficiency) and a smaller average distance increase (for representativeness). Following Rosenbaum's guideline,[14] we choose to use 25th percentile of the pairwise distances between TC and NTC patients, which leads to a template size of 2200. With this fixed template size, we draw 250 random samples from TC patients, and select the one that is closest to the full TC population in term of rank Mahalanobis distance as the final template sample. We utilize a logistic regression model to estimate the propensity score, which includes the eight covariates mentioned earlier, then conduct optimal matching with the estimated propensity score to pair the template sample with patients in NTC.

Figure 3 shows the balance of the eight covariates before and after matching in terms of the absolute standardized difference (ASD) in means. The covariate is considered well-balanced if the ASD is less than 10%. As shown in the figure, all ASD's are below 10% with most of them below 5%. We also calculate the variance ratio (VR) before and after matching for continuous covariates, age and ISS.[11] Both VR's are less than 1.5 after matching, indicating good balance. If we regard this well-matched data as coming from a randomized study, we can run a logistic regression model to estimate the risk difference.[20] The mortality difference is estimated to be 6.0% with a 95% confidence interval of [4.5%, 7.5%]. Therefore, barring unmeasured confounding, being admitted to TC causes the mortality rate to drop 6.0%, or equivalently a 6.0% increase of the survival rate, in the population of trauma patients who actually got admitted into TC.

4.2 | A further evaluation with assessment of hidden bias

The causal analysis in the previous subsection is based on the ignorable treatment assignment assumption. Assuming no unmeasured confounding, we regard the matched data as coming from a completely randomized experiment. A major limitation of such approach is the lack of assessment of unmeasured confounding, which is a big threat to the validity of observational studies. Like observed confounding, unmeasured confounding, if not adjusted for, can cause bias in causal effect estimation, which is often referred to as hidden bias.[4] Rosenbaum's sensitivity analysis assumes a hypothetical unmeasured confounder that associates with the treatment assignment. With a pre-specified magnitude of the association, bounds on the P-value can be identified for the causal effect hypothesis testing, which incorporates the maximal impact of the unmeasured confounder. By varying the magnitudes of the association, we can examine how the significance of the causal effect test changes. If the causal effect changes from statistically significant to insignificant just for a very small hidden bias, such observational study is regarded as being sensitive to unmeasured confounding. Rosenbaum's sensitivity analysis utilizes randomization tests based on the matched design, which is robust in the sense that no outcome modeling is involved.

Sensitivity analysis with continuous outcomes usually assumes a constant additive effect in testing.[21] This is not quite sensible for binary outcomes, as they can only take two distinct values. To conduct randomization test on binary outcomes with non-zero effect, Rosenbaum proposed a framework of attributable effect.[9] An effect is attributable to treatment if it

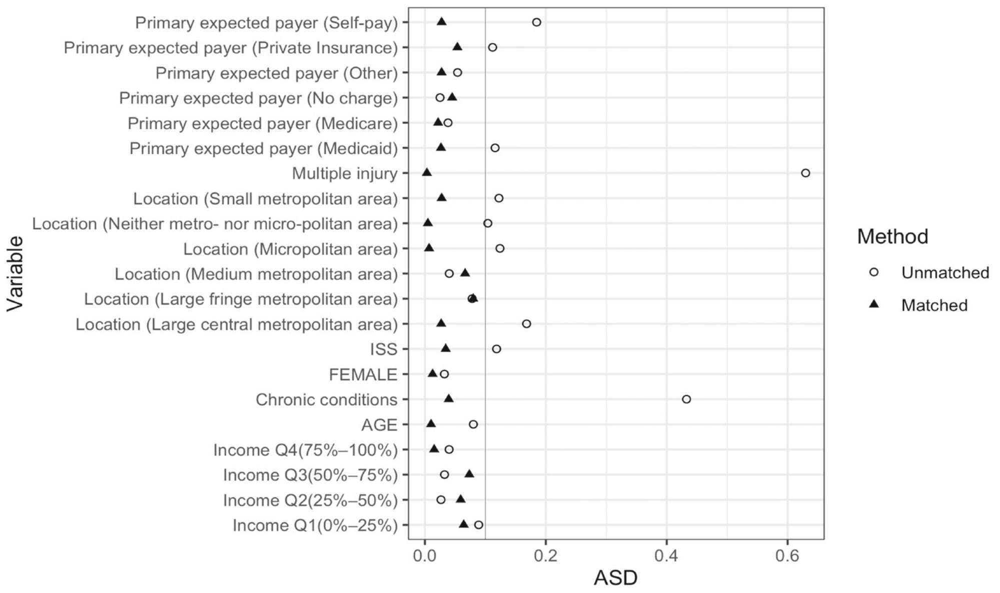

FIGURE 3 Variables balance before and after matching

would not have been observed had the individual not been exposed to treatment. Attributable effect can be calculated by inverting randomization test without distribution assumptions beyond the random assignment of treatments. Moreover, it is easy to incorporate sensitivity analysis in attributable effect estimation. For the rest of this section, we provide a brief review of the conceptual setup of attributable effect, followed by an application to our trauma care evaluation example.

4.2.1 | Attributable effect: A brief review

Under the potential outcome framework with binary data, the ith subject has two potential responses r_{Ti} and r_{Ci}, corresponding to the treatment status being treated and control. Assuming non-negative effect, that is, $r_{Ti} > r_{Ci}$, the potential response pair, (r_{Ti}, r_{Ci}), can take three possible values, $(1,1)$, $(1,0)$, or $(0,0)$. The treatment effect $\delta_i = r_{Ti} - r_{Ci}$ may be either 0 or 1, where 1 indicates the effect is attributable to receiving the treatment. The observed response is represented as $R_i = r_{Ti}Z_i + r_{Ci}(1 - Z_i) = r_{Ci} + Z_i\delta_i$, where Z_i is the treatment indicator. Then $A = \sum_{i=1}^{n} Z_i\delta_i$ is the number of events among treated subjects that are caused by the treatment (the number of lives saved because the patients are treated at TC), given a sample of size n. We cannot observe A since it involves potential outcomes, but we can observe $T = \sum_{i=1}^{n} Z_iR_i$, which is the number of events among treated subjects. To infer about the attributable effect, we consider a general null hypothesis of each individual effect, $H_0 : \boldsymbol{\delta} = \boldsymbol{\delta_0}$, where $\boldsymbol{\delta} = (\delta_1, \dots, \delta_n)^T$ and $\boldsymbol{\delta_0}$ is a vector of 0 and 1. Some of the null hypotheses are not logically possible given our non-negative effect assumption. A logically impossible hypothesis, also referred to as incompatible, can be rejected with type I error rate of zero. Among compatible hypotheses, we compute $A_0 = \sum_{i=1}^{n} Z_i\delta_{0i}$, which equals A if H_0 is true. Then a randomization test can be carried out based on a hypergeometric distribution and a confidence set can be identified via inverting the test.[22] A caveat is that, since A is defined only for the treated subjects, it carries a similar interpretation as the ATT effect.

Rosenbaum provided more details about how to estimate the attributable effect in matched design.[9] Suppose there are I matched sets and each matched set contains n_i subjects, indexed by $i = 1 \dots I$ and $j = 1 \dots n_i$. Using similar notations as above, $Z_{i+} = \sum_{j=1}^{n_i} Z_{ij}$ is number of treated subjects in set i. $R_{ij} = r_{Tij}Z_{ij} + r_{Cij}(1 - Z_{ij})$ is the observed response for a subject, and $R_{i+} = \sum_{j=1}^{n_i} R_{ij}$ is the number of subjects who have events in matched set i. $T = \sum_{i,j} Z_{ij}R_{ij}$ denotes the total number of treated individuals with events. $\delta_{ij} = r_{Tij} - r_{Cij}$ is the individual treatment effect, which is an unobserved variable. $r_{Ci+} = \sum_{j=1}^{n_i} r_{Cij}$ is the number of events if all subjects in a matched set were assigned to control. The attributable effect, A, and the number of treated subjects who would have events even in control, $T - A$, can be defined as below:

$$A = \sum_{i,j} Z_{ij}(r_{Tij} - r_{Cij}) = \sum_{i,j} Z_{ij}\delta_{ij}, \quad T - A = \sum_{i,j} Z_{ij}r_{Cij}. \tag{14}$$

Consider testing the null hypothesis $H_0 : \boldsymbol{\delta} = \boldsymbol{\delta_0}$ against the alternative $H_a : \boldsymbol{\delta} \geq \boldsymbol{\delta_0}, \boldsymbol{\delta_0} \neq \boldsymbol{\delta_0}$. Note that $\boldsymbol{\delta_0} = (\delta_{01}, \ldots, \delta_{0n})^T$ is a vector of pre-specified treatment effect for each individual. Therefore, it naturally allows for heterogeneous effects, as some of δ's are 1 and the others are 0.[9] If the null hypothesis is true, then r_{Cij} could be obtained for all subjects, because $r_{Cij} = R_{ij} - Z_{ij}\delta_{ij}$. To be more specific, $r_{Cij} = R_{ij}$ for a control subject, and for a treated subject, if the outcome is attributable to treatment then $r_{Cij} = R_{ij} - 1$, otherwise $r_{Cij} = R_{ij}$. Write $A_0 = \sum_{i,j} Z_{ij}\delta_{0ij}$, so that under null hypothesis $T - A_0 = \sum_{i,j} Z_{ij} r_{Cij}$, which is the sum of I independent binary random variables B_i, $i = 1, \ldots, I$, where $B_i = \sum_j Z_{ij} r_{Cij}$. Write $\pi_i = Pr(B_i = 1)$, under a randomized experiment, we know $\pi_i = \frac{Z_{i+} r_{Ci+}}{n_i}$. To test the null hypothesis against alternative hypothesis is to find the probability $Pr(\sum_i B_i \geq T - A_0)$, and as $I \to \infty$, $Pr(\sum_i B_i \geq T - A_0)$ can be approximated by

$$Pr\left(\sum_i B_i \geq T - A_0\right) \to 1 - \Phi\left(\frac{T - A_0 - \sum_i \pi_i}{\sqrt{\sum_i \pi_i(1 - \pi_i)}}\right), \tag{15}$$

where Φ is the CDF of standard normal distribution. In practice, if $Pr(\sum_i B_i \geq T - A_0) \geq 0.05$, we fail to reject the null hypothesis under 0.05 significance level and therefore $A_0 = \sum_{i,j} Z_{ij}\delta_{0ij}$ is a plausible attributable effect. Since $\boldsymbol{\delta_0}$ is a high dimensional vector of 1's and 0's, it is difficult to find the confidence interval for A_0 by going over each possible $\boldsymbol{\delta_0}$. It is more efficient to find one $\boldsymbol{\delta_0}$ with $A_0 = \sum_{i,j} Z_{ij}\delta_{0ij}$ that is hardest to reject. If this $\boldsymbol{\delta_0}$ is rejected, then we can conclude it is not plausible that A_0 or fewer events are attributed to treatment, otherwise A_0 is a plausible attributable effect.

4.2.2 | Attributable effect for trauma center patients

We apply the attributable effect analysis to our matched trauma care data (2200 pairs). For each patient, the response is either alive ($R_{ij} = 1$) or dead ($R_{ij} = 0$). We intend to examine, among patients treated at TC, how many of them were alive because they received the care at TC. This analysis is under the assumption that no patient would die if treated at TC, but would survive if treated at NTC. This is reasonable in the sense that TC has better resource and more clinical expertise and this should not have a harmful effect on patients. Table 5 reports the aliveness and death of patients by matched pairs. In most pairs, patients are both alive (1908) or both dead (17). In 204 discordant pairs, patients admitted to TC are alive and their matched patient admitted to NTC are dead. In another 71 discordant pairs, patients admitted to TC are dead and their matched patient admitted to NTC are alive. Among the 2200 pairs, $Z_{i+} = 1$ for every pair, $R_{i+} = 2$ for 1908 pairs, $R_{i+} = 1$ for 275 discordant pairs and $R_{i+} = 0$ for 17 pairs. Rosenbaum showed that in a cohort study the hardest to reject null hypothesis is assigning the attributable effect to concordant pairs ($R_{i+} = 2$).[9] Suppose we want to ask whether $A_0 = 102$ is a plausible attributable effect assuming no hidden bias. We assign all of them to concordant pairs and calculate π_i for each of the 2200 pairs. As a result, the 17 pairs with $R_{i+} = 0$ have $\pi_i = 0$, and the 275 discordant pairs with $R_{i+} = 1$ have $\pi_i = 1/2$. Among the 1908 concordant pairs, 1806 pairs have $\pi_i = 1$ and the other 102 pairs whose treatment effects are attributable have $\pi_i = 1/2$. Then the test statistic is

$$\frac{T - A_0 - \sum_i \pi_i}{\sqrt{\sum_i \pi_i(1 - \pi_i)}} = 1.60.$$

So the null hypothesis is plausible since 1.60 < 1.65 (critical value for one-sided z-test). In contrast, with $A_0 = 101$, $\frac{T - A_0 - \sum_i \pi_i}{\sqrt{\sum_i \pi_i(1-\pi_i)}} = 1.6502$, the null hypothesis is barely rejected. Thus, we can conclude under the assumption of no hidden biases, among the 2112 alive TC patients, at least 102 of them are attributable to being admitted to TC.

TABLE 5 Aliveness and death status for matched pairs

	NTC, dead	NTC, alive	NTC
TC, dead	17	71	88
TC, alive	204	1908	2112
TC	221	1979	2200

4.2.3 | Sensitivity analysis

The analysis in the previous subsection assumes no unmeasured confounding. The matched data are created based on the estimated propensity score from the eight observed covariates. Though they are all important covariates, it is likely that other factors may contribute to the underlying differences in TC and NTC patients. We further consider the sensitivity analysis to assess the robustness of the causal conclusion to departures from random assignment. Sensitivity analysis starts with the assumption that treatment assignment in a matched pair is strongly ignorable given the observed variables X and an additional unmeasured confounder U, which is for $i = 1, \dots, 2200, j = 1, 2$

$$(r_{Tij}, r_{Cij}) \perp\!\!\!\perp Z_{ij} | X_{ij}, U_{ij}, \quad 0 < Pr(Z_{ij} | X_{ij}, U_{ij}) < 1. \tag{16}$$

Denote $p_{ij} = Pr(Z_{ij} = 1 | X_{ij}, U_{ij})$, which is the probability of being assigned to treated group for patient j in matched pair i. Based on Rosenbaum's framework,[23,24] a sensitivity parameter Γ is introduced to capture the impact due to unmeasured confounding. Specifically, two subjects in a matched pair may differ in their odds of being exposed to treatment bounded by a function of Γ,

$$\frac{1}{\Gamma} \leq \frac{p_{i1}/(1 - p_{i1})}{p_{i2}/(1 - p_{i2})} \leq \Gamma. \tag{17}$$

The case of $\Gamma = 1$ yields random assignment of treatment which means $p_{ij} = 1/2$ and the unmeasured confounder does not exist. If $\Gamma > 1$, the randomization distribution of the statistic B_i is unknown since $\pi_i = Pr(B_i = 1)$ depends on the distribution of U_{ij}. In a cohort study, it is shown that π_i is bounded above by a quantity $\overline{\pi_i}$, $\overline{\pi_i} = \frac{\Gamma Z_{i+} r_{Ci+}}{(1-\Gamma) Z_{i+} r_{Ci+} + n_i}$. Thus, as $I \to \infty$, $Pr(\sum_i B_i \geq T - A_0)$ can be approximated by $1 - \Phi(\frac{T - A_0 - \sum_i \overline{\pi_i}}{\sqrt{\sum_i \overline{\pi_i}(1 - \overline{\pi_i})}})$. Then the hypothesis test can be carried out by replacing π_i with $\overline{\pi_i}$ to find the upper bound of the P-value. Technically, to conduct the sensitivity analysis, we first fix the value of parameter Γ, then compute $\overline{A_0}$, which is the smallest value of A_0 that makes the null hypothesis barely plausible under the α significance level. $\overline{A_0}$ is also the lower bound of the $(1 - \alpha) \times 100\%$ confidence set for all plausible attributable effects. A small value of $\overline{A_0}$ implies a more negligible attributable effect. If a small Γ (meaning a slight departure from random assignment) yields a small $\overline{A_0}$, the causal effect estimated with the observed data is considered sensitive to hidden bias. One may be interested in computing the smallest value of Γ that makes the attributable effect equal to a certain threshold. If the threshold is set to zero, the corresponding Γ value indicate the smallest association between the unmeasured confounder and the treatment that leads to no attributable effect. If a 1% aliveness rate difference between the two groups is considered meaningful, the threshold may be set to 1%.

Table 6 reports the estimates of the lower endpoint of 95% confidence set for the attributable effect ($\overline{AE}$) and the attributable risk ($\overline{AR}$, in percentage) as Γ increases. The attributable risk is the proportion of alive cases among TC patients, which can be attributed to being cared at TC. Without any hidden bias, at least 4.64% of alive cases are attributable to being treated at TC. How sensitive is this result to potential hidden biases? As Γ increases, both attributable effect and attributable risk decrease, which shows an increasing amount of the observed causal effect can be explained away by the unmeasured confounding. When $\Gamma = 1.5$, or one patient in any pair is 1.5 times more likely (in odds) than the other to be admitted to TC due to an unobserved factor, 42 or more alive patients in TC are caused by being treated at TC. This can also be interpreted as that the lives of at least 1.91% of all TC patients are saved by treating them at TC. The drop from 4.64% to 1.91% in the attributable risk is because we account for the unobserved factor with an impact of magnitude $\Gamma = 1.5$ in the hypothesis testing. The last row of the table has the attributable effect drop to zero, which implies that an unobserved factor with an impact of magnitude $\Gamma = 2.4$ could completely explain away any attributable effect in the data. Hence it is plausible that TC offers no benefit over NTC for trauma patients, if an unobserved factor makes the patients 2.4 times more likely (in odds ratio) to be sent to TC than to NTC. From a trauma care system evaluation perspective, if a 1% or more reduction in mortality is considered a worthwhile investment to keep TC, we may set the threshold for the attributable risk at 1%. The table shows that $\Gamma = 1.8$ results in $\overline{AR} = 1\%$. As long as the unmeasured confounder does not alter the odds ratio of being admitted to TC over NTC by more than 1.8, there is at least 1% of the TC patients whose lives are saved by being treated at TC.

As shown in supplementary material Section 3, we also assess the robustness of our causal analyses by varying the size of template sample. We consider two more scenarios with template size 2100 or 2300, and we find the results are very similar, which implies our findings are stable.

TABLE 6 Lower bound for attributable effect when hidden biases exist

Γ	$\overline{AE}$	$\overline{AR}$%
1	102	4.64
1.2	71	3.23
1.4	50	2.27
1.5	42	1.91
1.7	28	1.27
1.8	22	1.00
2.0	12	0.55
2.2	4	0.18
2.4	0	0.00

5 | DISCUSSION

We propose a flexible matched design to address challenges in the conventional matching framework for observational studies. These challenges include estimating ATT in a setting with a bigger treated group and a smaller control group, or estimating ATE with pair matched data. Our template matching first identifies a representative sample of the population of interest, then match subjects from the original dataset to this reference group. Theoretical justification shows unbiased treatment effect estimation is expected under regular assumptions, and simulation studies also demonstrate good empirical performance. With a real-world trauma care database, we use the matched design to control observed confounding and implement a sensitivity analysis to assess the impact due to unmeasured confounding.

Template matching may have a broader range of applications, since it only depends on identifying a representative template sample of the target population. In observational studies, researchers often rely on existing data sources to extract information about research questions of their interest, due to the lack of resource to design their own study and collect data. In health research, national surveys, patient registries or hospital patient records are commonly used. Many times, such existing databases may not align well with the ideal patient population that the researchers intend to study. Simply subsetting the existing dataset may not work, since the distribution of important patient characteristics in the subset may not match the target population. It would be desirable if a pseudo population can be created with pre-specified covariate distributions, such as age group distribution, gender composition, race/ethnicity distribution and so forth then matched to the original database to extract real patients from both exposed and unexposed groups to learn about causal effects in this special population. For example, the NEDS database, illustrated in Section 4, does not disclose geographical information. Researchers may get stuck if they want to learn about the impact of TC vs NTC in a particular state, say Ohio. The key characteristics of trauma patient in Ohio are often reported in literature and health agency's published documents. So researchers may use the public available information to construct a pseudo population with key characteristics similar to Ohio patients, then use it as a template to match to the NEDS database. This approach still requires more thoughts to be methodological sounding, such as all relevant covariates should be specified in creating template group and assumptions on joint distribution information. But more explorations are worthwhile as it opens the door to make better use of existing databases.

One potential limitation is the template sample selection. We suggest a numerical method that intends to minimize a certain distance metric. In practice, Matching methods are generally study/data dependent since the matching quality relies on the overlap of the covariate distribution. If the data overlap very well, the researchers can pick a template group easily by taking a random sample from the target population. But scenarios with moderate overlap, practitioners may get stuck on how to pick a sensible template sample size. We suggest trying different sizes and using an average distance based method to determine the template size. It may add a bit computational burden, but we believe this should help clinical researchers better implement template matching in practice. When the overlap between two groups is bad, matching alone may not work well, since high quality matches are hard to find and it relies heavily on extrapolation. Additional model based adjustment could be helpful if good knowledge about model specification is available.

ACKNOWLEDGEMENTS

This work was partially supported by Grant DMS-2015552 from National Science Foundation and Grant R01HD107280 from National Institute of Child Health & Human Development. The authors thank Junxin Shi and Henry Xiang for the dataset preparation and insightful discussion on comparative trauma care research. The authors also thank the associate editor and two anonymous reviewers for their insightful comments and suggestions.

DATA AVAILABILITY STATEMENT

We considered 5 years of data from the Nationwide Emergency Department Sample (NEDS), from 2006 to 2010. The NEDS is a national representative dataset designed by the Agency for Healthcare Research Quality to enable analyses of emergency department (ED) utilization patterns. We selected trauma patients using the National Trauma Data Standard Patient Inclusion Criteria as definition. These data were derived from the following resources: https://www.hcup-us.ahrq.gov/nedsoverview.jsp

The distribution of the NEDS database is regulated by the Healthcare Cost and Utilization project (HCUP) Central Distributor. The data are available for purchase and all the HCUP users must sign the Data Use Agreement form and complete an online training.

ORCID

Bo Lu https://orcid.org/0000-0002-3807-7869

REFERENCES

1. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika*. 1983;70(1):41-55.
2. Bang H, Robins JM. Doubly robust estimation in missing data and causal inference models. *Biometrics*. 2005;61(4):962-973.
3. Watkins S, Jonsson-Funk M, Brookhart MA, Rosenberg SA, O'Shea TM, Daniels J. An empirical comparison of tree-based methods for propensity score estimation. *Health Serv Res*. 2013;48(5):1798-1817.
4. Rosenbaum PR. *Design of Observational Studies*. New York: Springer; 2010.
5. Lu B. Causal inference for observational studies/real-world data. In: Yang H, Yu B, eds. *Real-World Evidence in Drug Development and Evaluation*. Boca Raton: Chapman and Hall/CRC; 2021:129-150.
6. Hansen BB. Full matching in an observational study of coaching for the SAT. *J Am Stat Assoc*. 2004;99(467):609-618.
7. Liu Y, Lu B, Foster R, et al. Matching design for augmenting the control arm of a randomized controlled trial using real-world data. *J Biopharm Stat*. 2022;32:124-140.
8. Bennett M, Vielma JP, Zubizarreta JR. Building representative matched samples with multi-valued treatments in large observational studies. *J Comput Graph Stat*. 2020;29(4):744-757.
9. Rosenbaum PR. Attributing effects to treatment in matched observational studies. *J Am Stat Assoc*. 2002;97(457):183-192.
10. Silber JH, Rosenbaum PR, Ross RN, et al. Template matching for auditing hospital cost and quality. *Health Serv Res*. 2014;49(5):1446-1474.
11. Imbens GW, Rubin DB. *Causal Inference in Statistics, Social, and Biomedical Sciences*. Cambridge: Cambridge University Press; 2015.
12. Nattino G, Lu B, Shi J, Lemeshow S, Xiang H. Triplet matching for estimating causal effects with three treatment arms: a comparative study of mortality by trauma center level. *J Am Stat Assoc*. 2021;116(533):44-53.
13. Nattino G, Song C, Lu B. Polymatching algorithm in observational studies with multiple treatment groups. *Comput Stat Data Anal*. 2022;167:107364.
14. Rosenbaum PR. Optimal matching of an optimally chosen subset in observational studies. *J Comput Graph Stat*. 2012;21(1):57-71.
15. Austin PC. A data-generation process for data with specified risk differences or numbers needed to treat. *Commun Stat Simul Comput*. 2010;39(3):563-577.
16. Sakran JV, Greer SE, Werlin E, McCunn M. Care of the injured worldwide: trauma still the neglected disease of modern society. *Scand J Trauma Resusc Emerg Med*. 2012;20(1):64.
17. AHRQ. Agency for healthcare research and quality, healthcare cost and utilization project. Introduction to the HCUP nationwide emergency department sample (NEDS); 2013. https://www.hcup-us.ahrq.gov/db/nation/neds/NEDS2013Introduction.pdf
18. Vickers BP, Shi J, Lu B, et al. Comparative study of ED mortality risk of US trauma patients treated at level I and level II vs nontrauma centers. *Am J Emerg Med*. 2015;33(9):1158-1165.
19. Shi J, Lu B, Wheeler KK, Xiang H. Unmeasured confounding in observational studies with multiple treatment arms: comparing emergency department mortality of severe trauma patients by trauma center level. *Epidemiology*. 2016;27(5):624-632.
20. Austin PC, Stuart EA. Estimating the effect of treatment on binary outcomes using full matching on the propensity score. *Stat Methods Med Res*. 2017;26(6):2505-2525.
21. Rosenbaum PR. Sensitivity analysis in observational studies. *Encycl Stat Behav Sci*. 2005;4:1809-1814.
22. Rosenbaum PR. Effects attributable to treatment: inference in experiments and observational studies with a discrete pivot. *Biometrika*. 2001;88(1):219-231.

23. Rosenbaum PR. Sensitivity analysis for certain permutation inferences in matched observational studies. *Biometrika*. 1987;74(1): 13-26.
24. Rosenbaum PR. Quantiles in nonrandom samples and observational studies. *J Am Stat Assoc*. 1995;90(432):1424-1431.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Zhao R, Lu B. Flexible template matching for observational study design. *Statistics in Medicine*. 2023;42(11):1760-1778. doi: 10.1002/sim.9698