RESEARCH ARTICLE

Matching algorithms for causal inference with multiple treatments

Anthony D. Scotina[1] [ID] | Roee Gutman[2]

[1]Department of Mathematics and Statistics, Simmons University, Boston, Massachusetts

[2]Department of Biostatistics, Brown University, Providence, Rhode Island

Correspondence
Anthony D. Scotina, Department of Mathematics and Statistics, Simmons University, Boston, MA 02115.
Email: anthony.scotina@simmons.edu

Funding information
Patient-Centered Outcomes Research Institute, Grant/Award Number: ME-1403-12104

Randomized clinical trials are ideal for estimating causal effects, because the distributions of background covariates are similar in expectation across treatment groups. When estimating causal effects using observational data, matching is a commonly used method to replicate the covariate balance achieved in a randomized clinical trial. Matching algorithms have a rich history dating back to the mid-1900s but have been used mostly to estimate causal effects between two treatment groups. When there are more than two treatments, estimating causal effects requires additional assumptions and techniques. We propose several novel matching algorithms that address the drawbacks of the current methods, and we use simulations to compare current and new methods. All of the methods display improved covariate balance in the matched sets relative to the prematched cohorts. In addition, we provide advice to investigators on which matching algorithms are preferred for different covariate distributions.

KEYWORDS
causal inference, generalized propensity score, matching, multiple treatments, observational data

1 | INTRODUCTION

1.1 | Overview

Multi-arm randomized clinical trials have been proposed as an efficient trial design to compare multiple treatments simultaneously. The efficiency stems from the fact that comparisons between multiple treatments can be observed in one trial, rather than performing several trials, each comparing only two treatments at a time.[1,2] This efficiency is more pronounced when multiple treatments are compared to a control group, where multiple control groups would be required when conducting each two-arm trial separately. However, multi-arm trials can be more complex to design, conduct, and analyze.[3] One design complication is that in multi-arm trials, all of the arms are required to follow a similar protocol with similar inclusion and exclusion criteria as well as primary and secondary outcomes. This is in contrast to multiple binary trials that may have different protocols for comparing each pair of treatments.[4] Another design challenge with multi-arm trials is to ensure that there is no bias in the comparison of multiple treatments, which can be introduced through imbalances in units' allocations.[4] A possible analysis complication is whether to make a statistical correction for the fact that multiple primary hypotheses are tested in the analysis.[5] Cook and Farewell[6] stated that when the examined hypotheses are pre-specified, represent orthogonal contrasts, and are analyzed separately, then it is reasonable to not adjust for multiplicity. On the other hand, Bender and Lange[7] argue that experiments with multiple arms should always adjust for multiplicity when multiple significance tests are used as primary hypotheses in a confirmatory setting.

These complications are exacerbated when comparing multiple treatments in nonrandomized settings, because those receiving one treatment may differ from those receiving another with respect to the covariates, **X**, which may also

influence the outcome.[8] The propensity score (PS), which is the probability of receiving treatment conditional on a set of observed covariates, is a common metric used in nonrandomized observational comparisons to adjust for observed differences in covariates between two treatment groups. Matching on the PS has been shown in theory[9-11] and in application[12,13] to be a design technique that can generate sets of units with similar covariates' distributions on average.

PS matching is commonly employed using a greedy nearest-neighbor algorithm that does not guarantee optimal balance among covariates.[14] Full matching was proposed as a possible algorithm to minimize the total sum of distances between matched units.[15] This algorithm is a special case of the minimum-cost flow problem and can be solved in polynomial time. Cardinality matching is another matching method that maximizes the cardinality or size of the matched sample subject to constraints on covariates' balance.[16] Cardinality matching solves a linear integer programming problem that may be computationally intensive.

For more than two treatment groups, the generalized propensity score (GPS) vector represents each unit's probability of receiving any one of the treatments, conditional on the covariates.[17,18] Comparison of units with similar GPS vectors provides unbiased unit-level estimates of the causal effects between multiple treatments. While the literature on matching for estimating causal effects of binary treatments is extensive,[16,19-23] generalizations and applications for multiple treatments remain limited, and currently, no optimal matching algorithm has been proposed. Moreover, optimally matching for multiple treatments, also known as k-dimensional matching, was shown to be an NP-hard problem,[24] and the computational complexity of such methods may make them impractical in problems with many units and multiple treatments.

In this paper, we propose several novel techniques for generating matched sets when estimating the causal effects between multiple treatments. We compare the performance of the new and previously proposed matching algorithms in balancing covariate distributions using simulation analyses. We also present a case study in which we compare the performance of different matching algorithms to estimate the effects of being released to different nursing homes from Rhode Island Hospital on the patient's propensity to be rehospitalized.

1.2 | Framework

For Z possible treatment groups, $W_i \in \mathcal{W} = \{1, \ldots, Z\}$ denotes the treatment group identification for unit $i \in \{1, \ldots, n\}$. Let n_w denote the size of treatment group w such that $\sum_{w=1}^{Z} n_w = n$, and T_{iw} is an indicator function that is equal to 1 if $W_i = w$ and to 0 otherwise. In addition, let $\mathcal{Y}_i = \{Y_i(1), \ldots, Y_i(Z)\}$ be the set of potential outcomes for unit i, where $Y_i(w)$ is the potential outcome for unit i if it were exposed to treatment w. In practice, only the potential outcome corresponding to the intervention that affected unit i is observed. The other potential outcomes cannot be observed because they correspond to treatment assignments that did not occur.[25,26] Assuming the Stable Unit Treatment Value Assumption,[27] the observed outcome for unit i can be written as

$$Y_i^{\text{obs}} = T_{i1} Y_i(1) + \cdots + T_{iZ} Y_i(Z).$$

Because we cannot directly observe the causal effect for unit i, we need to rely on multiple units of which some are exposed to each of the other $Z - 1$ possible treatments. For drawing causal inference, there are variables that are unaffected by W_i, ie, the covariates $\mathbf{X}_i = (X_{i1}, \ldots, X_{iP})$. A crucial piece of information that is needed for causal effect estimation is the assignment mechanism, which is the probability for each unit to receive one of the Z treatments. If the n units represent a random sample from an infinite super-population, then the assignment mechanism is individualistic,[8] and it is unconfounded if

$$P(W_i = w | \mathbf{X}_i, \mathcal{Y}_i, \phi) = P(W_i = w | \mathbf{X}_i, \phi) \equiv r(w, \mathbf{X}_i)$$

for all $i = 1, \ldots, n$,[8] where $r(w, \mathbf{X}_i)$ is referred to as the GPS for treatment w and unit i,[18] and ϕ is a vector parameter governing the distribution. The assignment mechanism is probabilistic if $0 < P(W = w | \mathbf{X}, \mathcal{Y})$ for all w. We assume that the assignment mechanism is probabilistic and unconfounded throughout, so that comparing individuals with similar $R(\mathbf{X}) \equiv (r(1, \mathbf{X}), \ldots, r(Z, \mathbf{X}))$ results in well-defined causal effects.

Commonly, $R(\mathbf{X}_i)$ is unknown, and only an estimate of it, $\hat{R}(\mathbf{X}_i)$, is available. Thus, the observed data for unit i comprise the vector $\{Y_i^{\text{obs}}, \mathbf{X}_i, W_i, \hat{R}(\mathbf{X}_i)\}$.

1.3 | Estimands

Causal effect estimands are functions of unit-level potential outcomes on a common set of units.[25,26] With Z treatments, possible estimates of interest are the $\binom{Z}{2}$ pairwise population average treatment effects (ATEs) between treatments j and k, $\tau_{jk} \equiv E(Y(j) - Y(k))$, for $(j, k) \in \mathcal{W}^2$ and $j \neq k$, where expectation is taken over the entire population. The pairwise

ATEs are transitive, such that $\tau_{ik} - \tau_{ij} = \tau_{jk} \ \forall \ i \neq j \neq k \in \mathcal{W}$. One possible extension of τ_{jk} would be to contrast treatments among a subset of units in the population receiving one of the Z treatments and obtain the population average treatment effect on the treated (ATT),[28] $\tau_{jk}^{t} \equiv E(Y(j) - Y(k)\,|\,W = t)$. For reference treatment group $W = t$, the ATTs are also transitive, such that $\tau_{ik}^{t} - \tau_{ij}^{t} = \tau_{jk}^{t}$, but transitivity of the ATTs does not extend to conditioning on different reference treatment groups. For example, $\tau_{ik}^{1} - \tau_{jk}^{2}$ is generally not equal to τ_{ij}^{1}.[8]

In observational studies, some units may have very low propensity to receive certain treatments. These units violate the positivity assumption of the assignment mechanism. One method to overcome this limitation is to restrict the analysis only to units that may receive all of the possible treatments. Lopez and Gutman[8] proposed to include in the analysis only units that are in the rectangular common support region $\hat{r}(w, \mathbf{X}) \in (\hat{r}_{\min}(w, \mathbf{X}), \hat{r}_{\max}(w, \mathbf{X})) \forall w \in \mathcal{W}$, where

$$\hat{r}_{\min}(w, \mathbf{X}) = \max\{\min(\hat{r}(w, \mathbf{X} \,|\, W = 1)), \dots, \min(\hat{r}(w, \mathbf{X} \,|\, W = Z))\}$$
$$\hat{r}_{\max}(w, \mathbf{X}) = \min\{\max(\hat{r}(w, \mathbf{X} \,|\, W = 1)), \dots, \max(\hat{r}(w, \mathbf{X} \,|\, W = Z))\}.$$

Letting E_i denote the indicator for all treatment eligibility, we have

$$E_i = \begin{cases} 1, & \text{if } \hat{r}(w, \mathbf{X}_i) \in (\hat{r}_{\min}(w, \mathbf{X}_i), \hat{r}_{\max}(w, \mathbf{X}_i)) \text{ for all } w \in \mathcal{W} \\ 0, & \text{otherwise.} \end{cases}$$

Then, ATTs for units within this rectangular region are defined as

$$\tau_{jk}^{t,E} = E\left(Y(j) - Y(k)\,|\,W = t, E_i = 1\right).$$

These estimands may restrict the analysis to a specific part of the population that may not be generalizable to the entire population who received treatment t. However, this restricted population is the one for whom all Z treatment have a positive probability of being administered and for which an equipoise between all Z treatments exists. In that sense, these estimands sacrifice some external validity to improve internal validity.[29] In many causal inference questions, having a credible and precise answer for a subpopulation is considered more important than having a controversial imprecise answer for the entire target population.[29]

The estimands τ_{jk} and τ_{jk}^{t} can be approximated using the sample ATEs, ie,

$$\hat{\tau}_{jk} \equiv \frac{1}{n} \sum_{i=1}^{n} (Y_i(j) - Y_i(k)), \quad \hat{\tau}_{jk}^{t} \equiv \frac{1}{n_t} \sum_{i=1}^{n} T_{it}(Y_i(j) - Y_i(k)).$$

These approximations are hypothetical, because one can only observe one of the potential outcomes for each unit. Matching procedures have been proposed as one possible solution to this problem.[8,30]

1.4 | Distance measures

To define the similarity between patients receiving different treatments, a distance measure is required. The linear GPS is one distance measure that can be used in the multiple-treatment setting, ie,

$$|\text{logit}(r(w, \mathbf{X}_i)) - \text{logit}(r(w, \mathbf{X}_j))|, \quad i \neq j \in \{1, \dots, n\}, \quad w \in \{1, \dots, Z\}.$$

Using the transformed GPS is preferred to a distance measure with the untransformed $r(w, \mathbf{X}_i)$ because matching on the logit transformation has been shown to produce lower bias in matched samples.[8,31,32] One limitation of this distance is that it only compares one component of the GPS at a time. Matching only on a single component may not necessarily result in similar values for all of the other components of the GPS vector. This limitation is exacerbated as the number of treatment groups increases.

The Euclidean distance is a multivariate matching metric that is the sum of the squared differences between the vector's components. Formally, the Euclidean distance between vectors $\mathbf{V}_i = (v_{i1}, \dots, v_{iP})$ and $\mathbf{V}_j = (v_{j1}, \dots, v_{jP})$ is

$$\left\{ (\mathbf{V}_i - \mathbf{V}_j)^{\mathrm{T}} (\mathbf{V}_i - \mathbf{V}_j) \right\}^{1/2} = \sqrt{\sum_{k=1}^{P} (v_{ik} - v_{jk})^2},$$

where P is the dimension of $\mathbf{V}_i$. A limitation of the Euclidean distance as a matching metric is its sensitivity to the correlation structure of the vector's components.[33,34]

This sensitivity to the correlation of covariates is mitigated with the Mahalanobis distance.[19,35,36] The Mahalanobis distance between $\mathbf{V}_i$ and $\mathbf{V}_j$ is

$$\left\{ \left(\mathbf{V}_i - \mathbf{V}_j\right)^{\mathrm{T}} \Sigma^{-1} \left(\mathbf{V}_i - \mathbf{V}_j\right) \right\}^{1/2},$$

where Σ is the covariance matrix of $\mathbf{V}_i$. Commonly, Σ is unknown, and an estimate of it, $\hat{\Sigma}$, is used instead. With binary treatment, matching on the Mahalanobis distance of the covariates performs well in reducing covariate bias between the two treatment groups when the covariates' space is small. However, the reduction in bias is less than optimal when the covariates are not normally distributed or there are many covariates.[37,38] When defining similarity of units with multiple treatments, the Euclidean and Mahalanobis distances can be applied to either the original covariates, the untransformed estimated GPS vector, or the transformed estimated GPS vector. Because the Mahalanobis distance is frequently preferred to the Euclidean distance, the latter distance was not examined in our simulation analysis.

2 | MATCHING ALGORITHMS

2.1 | Basic matching algorithms

Matching procedures for causal inference attempt to find units that are "close" to each other in terms of $\mathbf{X}$, but receive a different treatment. It is relatively straightforward to perform matching when only a single covariate influences the assignment to treatment, but this task can become more complex as the number of covariates increases.

The basic matching with replacement algorithm identifies for unit i, ie, the units with the shortest distances from each of the other treatment groups. Because this algorithm identifies matches to all of the units, some of the matches may not be very close in terms of the distance measure. One possible solution is to restrict all of the matches to have a distance that is smaller than a predefined threshold (caliper).

The basic matching algorithm with reference treatment $t \in \mathcal{W}$ is summarized as follows.

1. Estimate $R(\mathbf{X}_i)$, $i = 1, \ldots, n$, using a multinomial logistic regression model.
2. Drop units outside the common support (eg, those with $E_i = 0$) and refit the model once.
3. For all $t' \neq t$, match those receiving t to those receiving t' using a prespecified distance measure and a caliper.
4. Units receiving t who were matched to units receiving all treatments $w \neq t$, along with their matches receiving the other treatments, comprise the final matched cohort.

Yang et al[30] have implemented this procedure with the Mahalanobis distance of the covariates and the linear GPS in step 4, but with a slight modification to step 2. Instead of removing all units with $E_i = 0$, they proposed to keep only units for which $\sum_{w=1}^{Z} \frac{1}{r(w,\mathbf{X}_i)} \leq \lambda$, where λ is a predefined threshold parameter. Under the assumptions of constant treatment effects and homoscedasticity, threshold λ is estimated such that the variances of each pairwise treatment effect are minimized. A possible limitation of the trimming procedures is that units with different $R(\mathbf{X}_i)$ values may be dropped, resulting in dropped units with different covariates' distributions. Thus, we compared all of the methods only to units with $E_i = 1$.

2.2 | Vector matching

The basic matching algorithm relies on a distance measure that aggregates individual component differences over the entire vector. In some cases, it may result in some components of the matched vector that are far apart and other components that are relatively close. Vector matching (VM) is a possible algorithm for matching units in observational studies with multiple treatments, which address this limitation.[8] VM uses k-means clustering to group units such that within each cluster, units are roughly similar on the $Z - 2$ components of the GPS, and it performs the matching on the remaining components using the linear GPS only among units that are in the same cluster. The matching method proposed by Yang et al[30] is a special case of VM when $k = 1$. However, for $k > 1$, VM provides a valid estimate of the ATEs because it ensures that units which are matched to one another are nearly perfect on one component of the GPS and roughly similar on the other components.[8]

For a reference treatment $t \in \mathcal{W}$, the VM procedure is summarized as follows.

1. Estimate $R(\mathbf{X}_i)$, $i = 1, \ldots, n$, using a multinomial logistic regression model.
2. Drop units outside the common support (eg, those with $E_i = 0$) and refit the model once.

3. For all $t' \neq t$:

    (a) Partition all units using k-means clustering on $\text{logit}(\hat{R}_{t,t'}(\mathbf{X})) \equiv (\text{logit}(\hat{r}(w, \mathbf{X})) \ \forall \ w \neq t, t')$. This forms K strata of units with relatively similar components of the GPS vectors.

    (b) Within each strata $k \in \{1, \ldots, K\}$, match all units receiving t to those receiving t' on $\text{logit}(\hat{r}(t, \mathbf{X}_i))$ with replacement using a caliper of $\epsilon * \text{sd}(\text{logit}(\hat{r}(t, \mathbf{X}_i)))$, where $\epsilon = 0.5$.

4. Units receiving t who were matched to units receiving all treatments $w \neq t$, along with their matches, comprise the final matched cohort.

VM was shown to yield the lowest average maximum pairwise bias in comparison to other nearest-neighbor matching and weighting methods for multiple treatments, while retaining a high percentage of units matched.[8]

VM has several potential limitations. First, it is a 1:1:$\cdots$:1 nearest-neighbor matching algorithm. While this ensures that each unit will be paired with its best match, this could leave units outside the final matched cohort, resulting in larger sampling errors. However, the increase in sampling error is usually minimal as long as the reference group stays of the same size and only the compared groups decrease in size.[39,40] In addition, the gain in power may be hindered by increasing bias that arises from the introduction of poorer matches. Second, using an identical caliper for all units could result in significant loss of units in the reference group left without matches in the other $Z - 1$ groups. Third, when the number of treatments increases, it is harder for VM to ensure that the matched sets are balanced across all components of the GPS vector, because it only performs close matching on one component of $R(\mathbf{X})$. Fourth, because VM uses k-means clustering to classify units in step 3(a), possible matches that are on the boundaries of different clusters may not be considered, which may result in increased bias.

2.3 | Extensions to VM

We examine several variations of VM that attempt to address the limitations described in Section 2.2. To address the first limitation, we examined VM with two matches for each unit (eg, 1:2:2 matching for $Z = 3$) (VM2) and VM using matching without replacement in step 3(b) (VMnr). To address the second limitation, we examined omitting the caliper in step 3(b) (VMnc). To address the third limitation of VM, we examined using the Mahalanobis distance of $\{\text{logit}(\hat{r}(t, \mathbf{X})), \text{logit}(\hat{r}(t', \mathbf{X}))\}$ in step 3(b) with and without a caliper (KM and KMnc). To address the fourth limitation of VM, we used fuzzy clustering instead of k-means clustering in step 3(a) (VMF). To address limitations 2-4 simultaneously, we combined fuzzy clustering in step 3(a) with the Mahalanobis distance of $\{\text{logit}(\hat{r}(t, \mathbf{X})), \text{logit}(\hat{r}(t', \mathbf{X}))\}$ in step 3(b) with and without a caliper (FM and FMnc).

In contrast to k-means clustering, which assigns units to only one cluster, fuzzy clustering assigns each unit a probability to be in a cluster. A common fuzzy clustering algorithm is the fuzzy c-means clustering algorithm.[41] With K clusters, the c-means algorithm minimizes the generalized least squares error functional

$$J_m(U, v) = \sum_{i=1}^{n} \sum_{k=1}^{K} (u_{ik})^m \|\mathbf{V}_i - v_k\|^2,$$

where m is a predefined weighting component, $U = [u_{ik}]$ is a fuzzy c-partition of a matrix $\mathbf{V}$, and $v = (v_1, \ldots, v_K)$ is a vector of centers. Complete fuzzy clustering occurs when each unit has equal membership for all clusters (ie, each membership coefficient is $1/K$), and hard clustering occurs when each unit has a membership coefficient of 1 for one of the clusters.

We consider several procedures to match units when using fuzzy clustering. First, we modify VM such that we classify all units using fuzzy clustering on $\text{logit}(\hat{R}_{t,t'}(\mathbf{X}))$. Because fuzzy clustering does not explicitly assign a cluster to any unit, we assign a unit to cluster $k \in K$ if its membership coefficient for that cluster is at least $1/K$. Units assigned to multiple clusters can be matched to other units that share at least one cluster with them. Second, we examined matching on the Mahalanobis distance of $\{\text{logit}(\hat{r}(t, \mathbf{X})), \text{logit}(\hat{r}(t', \mathbf{X}))\}$, within fuzzy clusters of $\text{logit}(\hat{R}_{t,t'}(\mathbf{X}))$. Third, we examined the effect of a caliper on these matching procedures.

The fuzzy clustering procedures have similar steps 1, 2, and 4 as VM. Step 3 is modified to include fuzzy clustering based on the Mahalanobis distance. For a reference treatment $t \in \mathcal{W}$, these procedures can be summarized as follows.

1. Estimate $R(\mathbf{X}_i)$, $i = 1, \ldots, n$, using a multinomial logistic regression model.
2. Drop units outside the common support (eg, those with $E_i = 0$) and refit the model once.

TABLE 1 List of matching algorithms used

Label	Distance	Caliper (Y/N)	Clustering
VM[8]	linear GPS	Y	k-means
VM2	linear GPS	Y	k-means
VMnc	linear GPS	N	k-means
VMnr	linear GPS	Y	k-means
VMF	linear GPS	Y	fuzzy
KM	Mahalanobis logit($\hat{r}(t,\mathbf{X}), \hat{r}(t',\mathbf{X})$)	Y	k-means
KMnc	Mahalanobis logit($\hat{r}(t,\mathbf{X}), \hat{r}(t',\mathbf{X})$)	N	k-means
FM	Mahalanobis logit($\hat{r}(t,\mathbf{X}), \hat{r}(t',\mathbf{X})$)	Y	fuzzy
FMnc	Mahalanobis logit($\hat{r}(t,\mathbf{X}), \hat{r}(t',\mathbf{X})$)	N	fuzzy
LGPSM	Mahalanobis logit GPS vector	Y	NA
LGPSMnc	Mahalanobis logit GPS vector	N	NA
COVnc	Mahalanobis covariates	N	NA

3. For all $t' \neq t$:

   (a) Partition all units using fuzzy clustering on logit($\hat{R}_{t,t'}(\mathbf{X})$). A unit is part of cluster k if its membership coefficient is at least $1/K$.

   (b) Within each strata $k \in \{1, \dots, K\}$, match those receiving t to those receiving t' on either the linear GPS (VMF) or the Mahalanobis distance of $\{\text{logit}(\hat{r}(t,\mathbf{X})), \text{logit}(\hat{r}(t',\mathbf{X}))\}$ with replacement and a caliper of 0.5 standard deviations (FM). FMnc performs the matching on the Mahalanobis distance of $\{\text{logit}(\hat{r}(t,\mathbf{X})), \text{logit}(\hat{r}(t',\mathbf{X}))\}$ without a caliper.

4. Units receiving t who were matched to units receiving all treatments $w \neq t$, along with their matches receiving the other treatments, comprise the final matched cohort.

We summarize and explicate each of the matching algorithms in Table 1. All VM-related algorithms use the linear GPS as the distance measure, whereas clustering is performed using either k-means (VM, VM2, VMnc, VMnr) or fuzzy (VMF) clustering. The second class of algorithms used an initial clustering step using either k-means clustering (KM, KMnc) or fuzzy clustering (FM, FMnc) followed by matching on the Mahalanobis distance of $\{\text{logit}(\hat{r}(t,\mathbf{X})), \text{logit}(\hat{r}(t',\mathbf{X}))\}$ as the distance measure within clusters. The last class of methods, including LGPSM, LGPSMnc, and COVnc, includes basic matching algorithms that do not use any initial clustering.

3 | SIMULATIONS

3.1 | Evaluating balance of matched sets by simulation

To compare the performance of the different matching algorithms described in Section 2, we performed extensive simulations. Simulation configurations were either known to the investigator or can be estimated from the data. A P-dimensional $\mathbf{X}$ was generated for $n = n_1 + \cdots + n_Z$ units receiving one of $Z \in \{3, 5, 10\}$ treatments, $\mathcal{W} = \{1, \dots, Z\}$. For $Z = 3$, we generated sample sizes such that $n_2 = \gamma n_1$ and $n_3 = \gamma^2 n_1$. For $Z = 5$, we generated similar sample sizes for n_1, n_2, and n_3 as for $Z = 3$, and we set $n_4 = n_2$ and $n_5 = n_3$. For $Z = 10$, the treatment-group sample sizes for $n_1, \dots, n_5$ were the same as for $Z = 5$, and the sizes of treatment groups 6-10 were $n_{i+5} = n_i$, $i = 1, \dots, 5$.

The values of $\mathbf{X}$ were generated from multivariate skew t distributions such that

$$\mathbf{X}_i \mid \{W_i = w\} \sim \text{Skew-}t_{df,P}(\boldsymbol{\mu}_w, \boldsymbol{\Sigma}_w, \boldsymbol{\eta}).$$

For $Z \in \{3, 5\}$, $\boldsymbol{\mu}_w = \text{vec}(1_P \otimes b_w)$, where 1_P is a $P \times 1$ vector of 1's and b_w is the $Z \times 1$ vector such that the w value is equal to b and the rest are zeros. In addition, the covariance matrices $\boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2, \boldsymbol{\Sigma}_3, \boldsymbol{\Sigma}_4$, and $\boldsymbol{\Sigma}_5$ have respective diagonal entries of 1, $\sigma_2^2, \sigma_3^2, \sigma_2^2$, and σ_3^2, and λ elsewhere. For $Z = 10$, b_w is a 10×1 vector such that the w value is equal to b and the rest are zeros, and the covariate matrices $\boldsymbol{\Sigma}_w$, $w = 1, \dots, 10$ have respective diagonal entries of 1, $\sigma_2^2, \sigma_3^2, \sigma_2^2, \sigma_3^2, 1, \sigma_2^2, \sigma_3^2, \sigma_2^2$, and σ_3^2, and λ elsewhere. This was done in order to reduce the running time when dealing with a large number of treatments.

TABLE 2 Simulation factors

Factor	$Z \in \{3,5\}$ Levels	$Z = 10$ Levels
n_1	$\{600, 1200\}$	900
γ	$\{1, 2\}$	$\{1, 2\}$
b	$\{0, 0.25, 0.50, 0.75, 1\}$	$\{0, 0.25, 0.50, 0.75, 1\}$
λ	$\{0, 0.25\}$	$\{0, 0.25\}$
σ_2^2	$\{0.5, 1, 2\}$	$\{1, 2\}$
σ_3^2	$\{0.5, 1, 2\}$	$\{1, 2\}$
η	$\{-3.5, 0, 3.5\}$	$\{0, 3.5\}$
df	$\{7, \infty\}$	$\{7, \infty\}$
P	$\{5, 10, 20\}$	$\{10, 20\}$

The simulation design assumes a regular assignment mechanism that depends on the factors listed in Table 2, resulting in a $2^4 \times 3^4 \times 5$ factorial design for $Z \in \{3, 5\}$ and a $2^7 \times 5$ factorial design for $Z = 10$. For each configuration, 100 replications were produced. For $Z \in \{3, 5\}$, we discarded configurations when $P = 20$ and $n_1 = 600$, and for $Z \in \{3, 5, 10\}$, we discarded configurations when $P = 20$ and $b = 1$, because of the small number of units that can be matched across all treatment arms. After discarding these configurations, there were 5184 simulation configurations for $Z \in \{3, 5\}$ and 576 simulation configurations for $Z = 10$. The simulation design is an extension to the ones described in the work of Lopez and Gutman.[8] The main differences between the simulation designs are the inclusion of more covariates as well as more possible interventions.

To estimate the GPS model, we used two different multinomial logistic regression models. The first model includes all of the covariates, $\mathbf{X}_i$, as explanatory variables. The second model includes all of the covariates as well as the square of all covariates, $\mathbf{X}_i^2$. Throughout, we will refer to the first model as the first-order GPS model and to the second model as the second-order GPS model. All of the simulations were conducted using R Studio software,[42] and all matching algorithms were implemented using the Matching package.[43]

3.2 | Simulation metrics

Let ψ_{iw} be the number of times that unit i serves as a match to other units in treatment group w, and let n_{wm} be the number of units from treatment group w in the matched sample, including units that are used as a match more than once. The weighted mean of covariate p, $p = 1, \dots, P$, at treatment w, is defined as $\bar{X}_{pw}$, such that

$$\bar{X}_{pw} = \frac{1}{n_{wm}} \sum_{i=1}^{n} X_{pi} T_{iw} \psi_{iw}.$$

We define the standardized bias at each covariate p for each pair of treatments as

$$SB_{pjk} = \frac{\bar{X}_{pj} - \bar{X}_{pk}}{\delta_{pt}},$$

where δ_{pt} is the standard deviation of $\mathbf{X}_p$ in the full sample among units receiving the reference treatment $W = t$. As in the works of McCaffrey et al[28] and Lopez and Gutman,[8] we extract the maximum absolute standardized pairwise bias at each covariate, as follows:

$$\text{Max2SB}_p = \max \left(|SB_{p12}|, |SB_{p13}|, |SB_{p23}|, \dots \right).$$

This metric reflects the largest discrepancy in estimated means between any two treatment groups for covariate p. McCaffrey et al[28] advocated a cutoff of 0.20 but maintained that larger cutoffs may be appropriate for different applications.

For matching algorithms with a caliper, a second metric that is used to measure matching performance is the proportion of units from the eligible population with $W = 1$ that were included in the final matched set, ie, *Prop.Matched*. When estimating τ_{jk}^t, scenarios with *Prop.Matched* ≈ 1 and low *Max2SB$_p$* for all p are optimal, because most of the units in the population of interest are retained, and the covariates' distributions are similar, on average, across treatment groups. By design, matching algorithms without a caliper have *Prop.Matched* $= 1$.

A third metric that is used to measure balance among matched sets is the ratio of the variances of covariates between each treatment group.[38,44] In the multiple-treatment setting, we extract the maximum absolute pairwise difference

between variances at each covariate, measured on the natural logarithm scale, ie,

$$\text{Max2log}_p = \max \left(\left| \log \left(\frac{\text{Var}(X_{p1})}{\text{Var}(X_{p2})} \right) \right|, \left| \log \left(\frac{\text{Var}(X_{p1})}{\text{Var}(X_{p3})} \right) \right|, \left| \log \left(\frac{\text{Var}(X_{p2})}{\text{Var}(X_{p3})} \right) \right|, \ldots \right),$$

where

$$\text{Var}(X_{pw}) = \frac{\sum_{i=1}^{n} (X_{pi} - \bar{X}_{pw})^2 T_{iw} \psi_{iw}}{n_{wm}}.$$

This metric summarizes the balance of the second moments of covariate p among all treatment groups. For binary treatment, Rubin[44] classified the ratio of variances between treatment groups that is in $[4/5, 5/4]$ as "good," the variance ratios in $[1/2, 4/5)$ or $(5/4, 2]$ as "of concern," and the variance ratios below $1/2$ or above 2 as "bad." These thresholds correspond to cutoffs of 0.22 and 0.69 on the absolute log scale.

For each configuration, at each replication, we calculated Max2SB_p and Max2log_p for all p, and *Prop.Matched*. We summarized Max2SB_p for all P covariates using $\text{MaxMax2SB} \equiv \max_{p=1,\ldots,P}(\text{Max2SB}_p)$ and $\overline{\text{Max2SB}} \equiv (1/P) \sum_{p=1}^{P} \text{Max2SB}_p$, and we summarized Max2log_p for all P covariates using $\text{MaxMax2log} \equiv \max_{p=1,\ldots,P}(\text{Max2log}_p)$. Lastly, we averaged MaxMax2SB, $\overline{\text{Max2SB}}$, and MaxMax2log across the 100 replications.

We provide results for MaxMax2SB because it represents the largest imbalance among all covariates, and it is not as affected as $\overline{\text{Max2SB}}$ by increasing the number of covariates with no biases. The results for $\overline{\text{Max2SB}}$ show similar trends and are available in the Appendix.

3.3 | Simulation results, $Z = 3$

Figure 1 shows boxplots of MaxMax2SB across simulation factors for $Z = 3$, for the prematched cohort of eligible units and the 10 matching algorithms, when using a first-order GPS model. Each point in the boxplots represents the maximal bias at one factors' configuration. The best-performing method is VMnr with only 12% of the configurations above 0.2. It is followed by LGPSM and FM, with 17% and 18% of the configurations above 0.2, respectively. COVnc has the worst performance, with MaxMax2SB exceeding 0.2 in 38% of configurations.
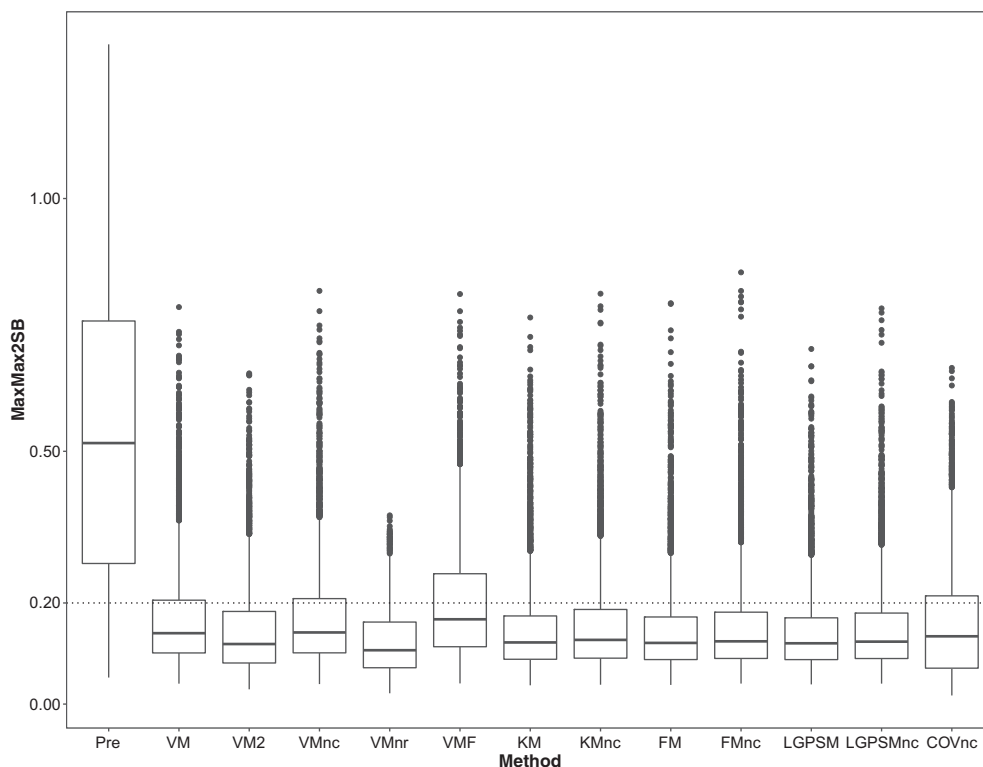

FIGURE 1 MaxMax2SB for a prematched cohort and by matching algorithms, $Z = 3$
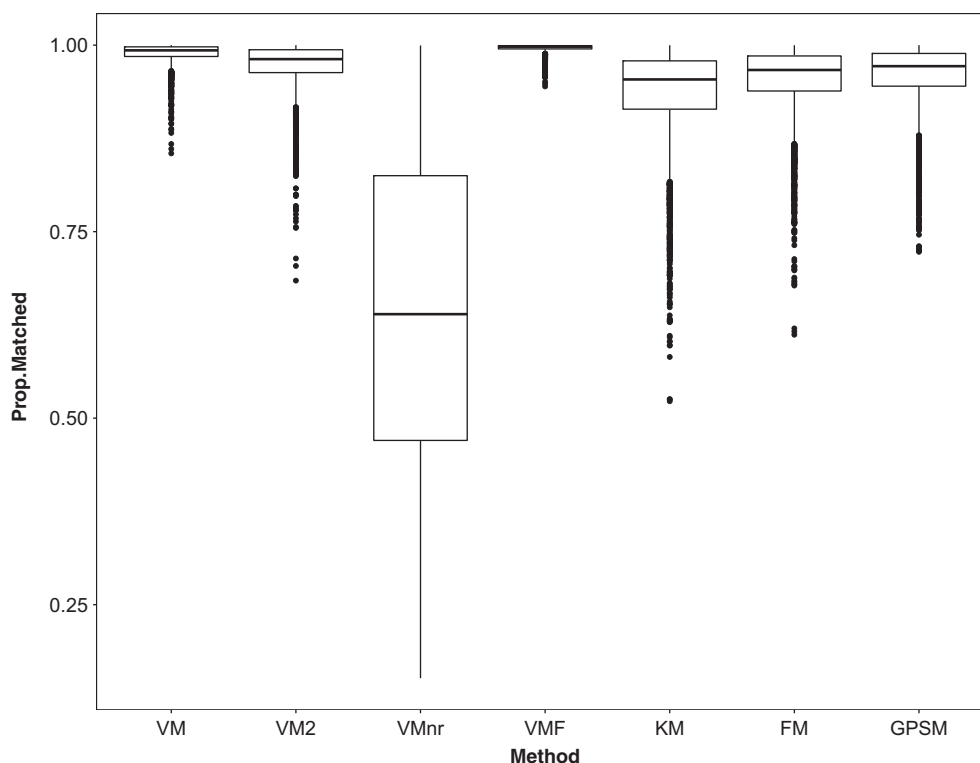
FIGURE 2 *Prop.Matched* for algorithms that used a caliper, $Z = 3$

When using a second-order GPS model, the best-performing method with second-order GPS is VMnr with 38% of the configurations above 0.2. It is followed by COVnc, LGPSM, and FM, with MaxMax2*SB* exceeding 0.2 in 39%, 45%, and 45% of configurations, respectively.

Figure 2 depicts *Prop.Matched* across configurations for matching algorithms with a caliper. VMF has the highest median *Prop.Matched* with 99.8% of the reference group units being matched on average. VM and VM2 have the second and third highest median *Prop.Matched*, with 99.3% and 98.1% of the reference group units matched on average, respectively. VMnr has the lowest median *Prop.Matched*, with only 64% of the reference group units matched on average. Thus, while VMnr generally yields the lowest bias in the matched cohort, it is at the expense of generalizability, because the matched cohort is less representative of the original sample.

To identify factors with the largest influence on the performance of each matching algorithm, we rank them by their MSE for MaxMax2*SB* (as in the works of Lopez and Gutman,[8] Rubin,[35] and Cangul et al[45]). For a first-order GPS model, initial covariate bias, b, explains the largest portion of variation in MaxMax2*SB*, accounting for at least 35% of the variability for each algorithm (data not shown). The number of covariates, P, explains the second-largest portion of variation in MaxMax2*SB*, accounting for at least 20% of the variability for most of the procedures, except for VMnr. Other influential factors include the interaction between b and P, and the ratio of units with $W = 2$ to units with $W = 1$ (γ), although these do not account for as much variability in MaxMax2*SB* as b or P. Similar trends are observed for the second-order GPS (data not shown).

Table 3 shows MaxMax2*SB* based on different levels of b, P, and the two GPS models. In settings with low covariate bias, all 10 matching algorithms appear to balance the covariates properly, regardless of the number of total covariates or GPS model order. As b increases, MaxMax2*SB* increases for all algorithms, although it increases more rapidly for matching on the Mahalanobis distance of the covariates (COVnc) when $P \geq 10$. For $P = 5$, MaxMax2*SB* increases at roughly the same rate for all matching algorithms as b increases. Except for COVnc, the median MaxMax2*SB* for each algorithm is higher across b and P when using a second-order GPS model. Among the methods that retain all of the units in $W = 1$, LGPSMnc has the lowest overall bias for most of the configurations, and it is comparable to VMnr.

The MaxMax2log for sets obtained using either VM, FMnc, LGPSMnc, or COVnc is depicted in Figure 3. For each algorithm, the median MaxMax2log is significantly larger than the 0.22 cutoff when using the first-order GPS. When including the square of the covariate terms in the GPS model, the MaxMax2log is lower for each algorithm that uses the

TABLE 3 Median MaxMax2*SB* across levels of *b*, *P*, and generalized propensity score (GPS) model order, $Z = 3$ (smallest MaxMax2*SB* in italics)

First-Order GPS Model												
$P=5$												
b	VM	VM2	VMnc	VMnr	VMF	KM	KMnc	FM	FMnc	LGPSM	LGPSMnc	COVnc
0.00	0.09	0.07	0.09	0.06	0.09	0.08	0.08	0.08	0.08	0.08	0.08	*0.04*
0.25	0.11	0.09	0.11	0.08	0.11	0.10	0.10	0.10	0.10	0.10	0.10	*0.06*
0.50	0.13	0.11	0.13	0.11	0.13	0.11	0.12	0.11	0.11	0.11	0.11	*0.10*
0.75	0.17	0.15	0.17	0.15	0.16	*0.13*	0.14	*0.13*	*0.13*	*0.13*	*0.13*	0.15
1.00	0.22	0.20	0.23	0.21	0.21	0.17	0.18	*0.16*	0.17	*0.16*	0.17	0.21
$P=10$												
b	VM	VM2	VMnc	VMnr	VMF	KM	KMnc	FM	FMnc	LGPSM	LGPSMnc	COVnc
0.00	0.08	0.06	0.08	*0.05*	0.08	0.08	0.08	0.07	0.07	0.07	0.07	0.07
0.25	0.09	0.07	0.09	*0.06*	0.09	0.09	0.09	0.09	0.09	0.09	0.09	0.13
0.50	0.12	0.10	0.12	*0.09*	0.12	0.11	0.11	0.11	0.11	0.11	0.11	0.21
0.75	0.18	0.16	0.18	*0.13*	0.17	0.15	0.16	0.15	0.16	0.15	0.16	0.31
1.00	0.27	0.24	0.27	*0.18*	0.26	0.24	0.25	0.23	0.24	0.22	0.23	0.41
$P=20$												
b	VM	VM2	VMnc	VMnr	VMF	KM	KMnc	FM	FMnc	LGPSM	LGPSMnc	COVnc
0.00	0.13	0.09	0.12	*0.08*	0.12	0.12	0.12	0.11	0.12	0.11	0.12	0.10
0.25	0.15	0.12	0.15	*0.10*	0.15	0.14	0.14	0.14	0.14	0.13	0.14	0.19
0.50	0.24	0.21	0.25	*0.16*	0.24	0.22	0.24	0.21	0.23	0.21	0.22	0.32
0.75	0.43	0.38	0.44	*0.26*	0.43	0.39	0.42	0.39	0.41	0.37	0.40	0.46
Second-Order GPS Model												
$P=5$												
b	VM	VM2	VMnc	VMnr	VMF	KM	KMnc	FM	FMnc	LGPSM	LGPSMnc	COVnc
0.00	0.12	0.10	0.13	0.09	0.14	0.11	0.11	0.11	0.11	0.11	0.11	*0.04*
0.25	0.13	0.11	0.13	0.11	0.17	0.11	0.12	0.11	0.12	0.11	0.12	*0.06*
0.50	0.15	0.13	0.15	0.13	0.20	0.12	0.13	0.12	0.13	0.12	0.12	*0.10*
0.75	0.18	0.17	0.19	0.18	0.26	0.15	0.16	*0.14*	0.15	0.15	0.15	0.15
1.00	0.24	0.23	0.24	0.24	0.34	0.19	0.21	*0.18*	0.20	*0.18*	0.19	0.21
$P=10$												
b	VM	VM2	VMnc	VMnr	VMF	KM	KMnc	FM	FMnc	LGPSM	LGPSMnc	COVnc
0.00	0.16	0.14	0.17	0.12	0.18	0.15	0.16	0.15	0.15	0.15	0.15	*0.07*
0.25	0.18	0.15	0.18	0.14	0.20	0.16	0.17	0.16	0.17	0.16	0.17	*0.13*
0.50	0.23	0.21	0.24	*0.17*	0.26	0.21	0.22	0.21	0.22	0.20	0.21	0.22
0.75	0.33	0.30	0.34	*0.24*	0.36	0.29	0.31	0.29	0.30	0.27	0.29	0.32
1.00	0.47	0.42	0.49	*0.32*	0.50	0.42	0.46	0.41	0.44	0.39	0.43	0.42
$P=20$												
b	VM	VM2	VMnc	VMnr	VMF	KM	KMnc	FM	FMnc	LGPSM	LGPSMnc	COVnc
0.00	0.19	0.16	0.19	0.14	0.19	0.19	0.19	0.19	0.19	0.18	0.19	*0.10*
0.25	0.23	0.21	0.24	*0.15*	0.25	0.22	0.23	0.22	0.23	0.22	0.22	0.19
0.50	0.36	0.32	0.36	*0.21*	0.37	0.34	0.36	0.34	0.35	0.33	0.34	0.32
0.75	0.57	0.50	0.58	*0.31*	0.57	0.53	0.58	0.53	0.57	0.50	0.54	0.46

GPS for matching. VM, FMnc, and LGPSMnc yield MaxMax2log below 0.22 for 24%, 34%, and 34% of configurations and below 0.69 for 93%, 92%, and 93% of configurations, respectively.

The most influential simulation factors for MaxMax2log were σ_2^2 and σ_3^2 for both GPS models. Table 4 shows the median MaxMax2log based on different levels of σ_2^2, σ_3^2, and the two GPS models. For VM, FMnc, and LGPSMnc, MaxMax2log is smallest when $\sigma_2^2 = \sigma_3^2 = 1$. This configuration implies that the initial variances are equal for all three treatment groups across all of the covariates. MaxMax2log is largest when the distance between σ_2^2 and σ_3^2 is largest, such as when $\sigma_2^2 = 2.0$ and $\sigma_3^2 = 0.5$, or vice versa. These trends are observed for both GPS models, although MaxMax2log was smaller overall for each level of σ_2^2 and σ_3^2 when using a second-order GPS model. Because COVnc only uses the GPS model in defining the region of common support, it is only slightly impacted by the choice of the GPS model. Thus, the smaller bias in the covariates when using COVnc is at the expense of larger variance ratios.
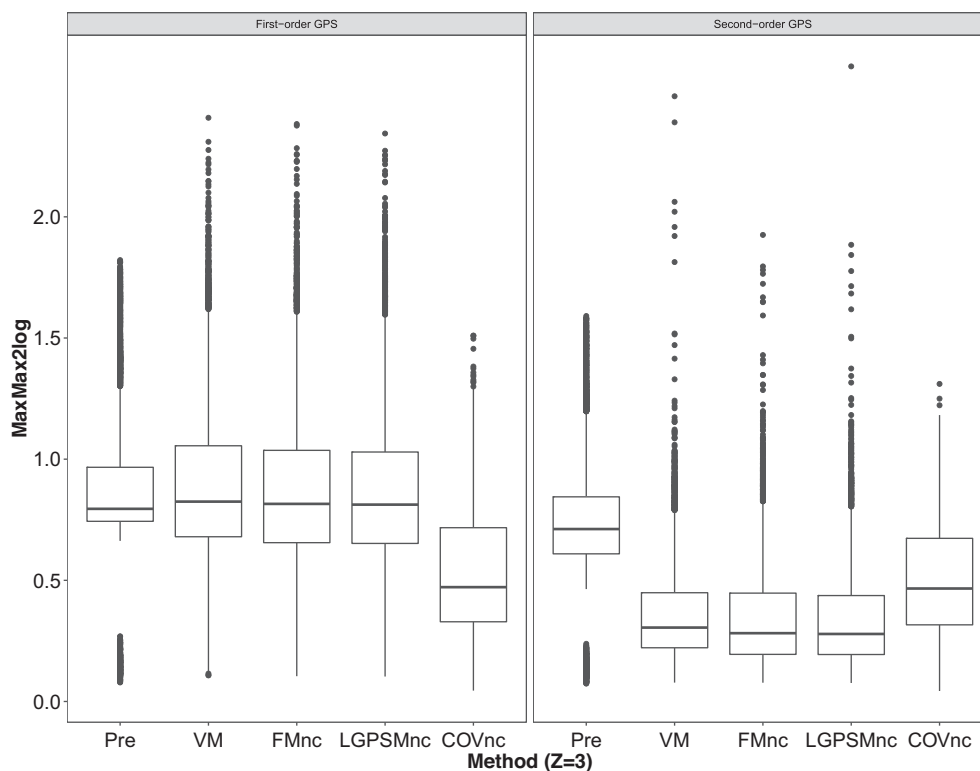
FIGURE 3 MaxMax2log for a prematched cohort across matching algorithms, $Z = 3$

TABLE 4 Median MaxMax2log across levels of σ_2^2, σ_3^2, and generalized propensity score (GPS) model order, $Z = 3$

	σ_2^2	σ_3^2	VM	FMnc	LGPSMnc	COVnc
	0.5	0.5	0.77	0.76	0.77	0.67
	0.5	1.0	0.78	0.77	0.77	0.66
	0.5	2.0	1.40	1.38	1.38	0.70
	1.0	0.5	0.77	0.77	0.76	0.63
First-order GPS	1.0	1.0	0.27	0.25	0.25	0.34
	1.0	2.0	0.81	0.80	0.80	0.38
	2.0	0.5	1.39	1.36	1.36	0.66
	2.0	1.0	0.81	0.80	0.80	0.38
	2.0	2.0	0.81	0.80	0.80	0.20
	0.5	0.5	0.37	0.38	0.37	0.61
	0.5	1.0	0.34	0.34	0.33	0.61
	0.5	2.0	0.37	0.33	0.32	0.63
	1.0	0.5	0.33	0.32	0.32	0.59
Second-order GPS	1.0	1.0	0.22	0.20	0.20	0.34
	1.0	2.0	0.28	0.23	0.23	0.38
	2.0	0.5	0.35	0.31	0.30	0.61
	2.0	1.0	0.28	0.24	0.24	0.38
	2.0	2.0	0.23	0.22	0.22	0.20

3.4 | Simulation results, $Z = 5$

Figure 4 shows boxplots of MaxMax2*SB* across simulation factors for $Z = 5$, for each of the 10 matching algorithms and in the prematched cohort of eligible units. VM exceeds 0.2 for the majority of the configurations. The algorithms
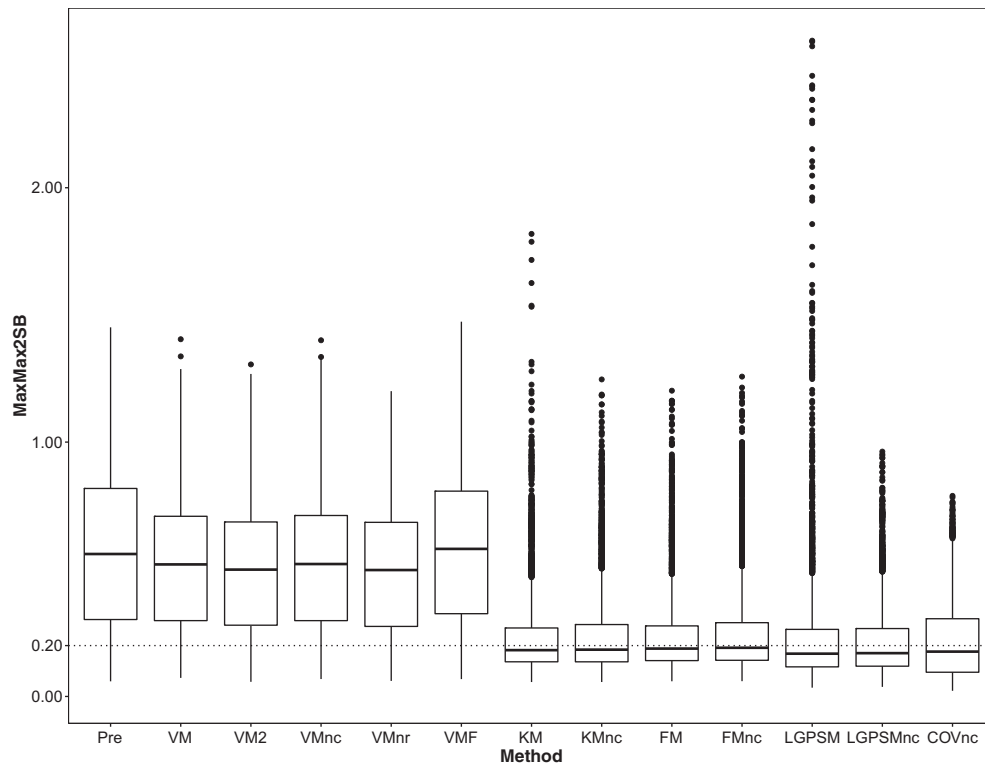
FIGURE 4 MaxMax2*SB* for a prematched cohort and by matching algorithms, $Z = 5$
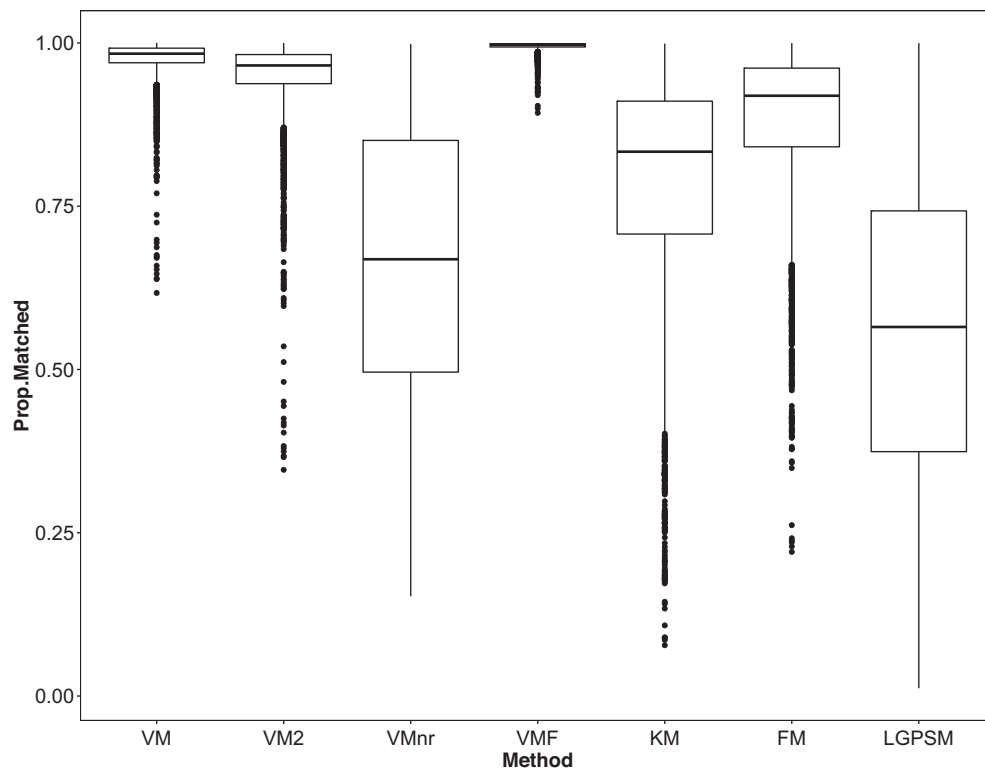

FIGURE 5 *Prop.Matched* for algorithms that used a caliper, $Z = 5$

without a caliper (KMnc, FMnc, LGPSMnc, COVnc) perform more favorably for five treatments, with the majority of the configurations yielding a MaxMax2*SB* below 0.20. KM and FM also perform favorably, with 57% and 55% of configurations yielding a MaxMax2*SB* below 0.20, respectively.

Similar to $Z = 3$, when using a second-order GPS model, the best-performing methods are similar to those when using a first-order GPS model, although MaxMax2SB is generally higher (data not shown). The algorithms without a caliper, which use a first-order GPS model (KMnc, FMnc, LGPSMnc), yield a MaxMax2SB below 0.20 for 41%, 41%, and 44% of configurations, respectively. KM and FM perform similarly, with 47% and 45% of configurations yielding a MaxMax2SB below 0.20, respectively. COVnc performs the best when using a second-order GPS model, with 56% of configurations yielding a MaxMax2SB below 0.20.

The only matching algorithms with a caliper that yield a median *Prop.Matched* exceeding 0.90 are the VM with replacement algorithms and FM (Figure 5). FM matches more than 75% of reference group units in 68% of configurations, and VM with replacement algorithms match more than 75% of the units in the reference group in 98% of the configurations.

TABLE 5 Median MaxMax2SB across levels of b, P, and generalized propensity score (GPS) model order, $Z = 5$ (smallest MaxMax2SB in italics)

|					**First-Order GPS Model**							
						P = 5						
b	VM	VM2	VMnc	VMnr	VMF	KM	KMnc	FM	FMnc	LGPSM	LGPSMnc	COVnc
0.00	0.14	0.12	0.14	0.11	0.14	0.11	0.10	0.11	0.11	0.08	0.08	*0.04*
0.25	0.33	0.31	0.33	0.31	0.33	0.13	0.13	0.14	0.14	0.10	0.10	*0.07*
0.50	0.54	0.53	0.54	0.53	0.56	0.15	0.15	0.17	0.17	*0.12*	0.13	*0.12*
0.75	0.70	0.69	0.70	0.71	0.73	0.18	0.19	0.20	0.21	*0.14*	0.18	0.17
1.00	0.82	0.81	0.83	0.86	0.86	0.24	0.25	0.25	0.26	*0.18*	0.24	0.24
						P = 10						
b	VM	VM2	VMnc	VMnr	VMF	KM	KMnc	FM	FMnc	LGPSM	LGPSMnc	COVnc
0.00	0.17	0.13	0.17	0.13	0.16	0.14	0.14	0.14	0.14	0.14	0.12	*0.08*
0.25	0.35	0.33	0.35	0.32	0.36	0.16	0.16	0.17	0.16	0.17	*0.15*	*0.15*
0.50	0.54	0.52	0.54	0.52	0.56	0.21	0.21	0.21	0.21	0.21	*0.20*	0.24
0.75	0.69	0.66	0.70	0.67	0.71	*0.29*	0.31	0.30	0.31	0.30	*0.29*	0.35
1.00	0.85	0.81	0.86	0.79	0.84	*0.44*	0.48	0.47	0.50	0.45	*0.44*	0.47
						P = 20						
b	VM	VM2	VMnc	VMnr	VMF	KM	KMnc	FM	FMnc	LGPSM	LGPSMnc	COVnc
0.00	0.16	0.13	0.16	0.12	0.16	0.14	0.14	0.14	0.14	0.14	0.13	*0.11*
0.25	0.35	0.33	0.35	0.32	0.36	0.17	0.17	0.17	0.18	0.17	*0.16*	0.21
0.50	0.53	0.51	0.54	0.48	0.54	0.27	0.29	0.28	0.30	*0.26*	*0.26*	0.35
0.75	0.74	0.69	0.75	0.62	0.73	0.52	0.56	0.53	0.60	*0.48*	*0.48*	0.50
					Second-Order GPS Model							
						P = 5						
b	VM	VM2	VMnc	VMnr	VMF	KM	KMnc	FM	FMnc	LGPSM	LGPSMnc	COVnc
0.00	0.17	0.15	0.18	0.14	0.19	0.14	0.15	0.14	0.15	0.13	0.14	*0.04*
0.25	0.37	0.35	0.37	0.35	0.38	0.13	0.15	0.14	0.15	0.11	0.13	*0.07*
0.50	0.55	0.53	0.55	0.54	0.60	0.15	0.16	0.15	0.16	*0.12*	0.15	*0.12*
0.75	0.70	0.69	0.70	0.70	0.76	0.17	0.20	0.18	0.19	*0.15*	0.20	0.17
1.00	0.83	0.82	0.83	0.86	0.90	0.22	0.26	0.23	0.26	*0.19*	0.27	0.24
						P = 10						
b	VM	VM2	VMnc	VMnr	VMF	KM	KMnc	FM	FMnc	LGPSM	LGPSMnc	COVnc
0.00	0.21	0.18	0.21	0.17	0.22	0.18	0.18	0.18	0.18	0.17	0.17	*0.08*
0.25	0.38	0.36	0.39	0.35	0.40	0.19	0.20	0.20	0.20	0.18	0.18	*0.15*
0.50	0.57	0.54	0.57	0.56	0.62	0.24	0.25	0.24	0.25	0.24	*0.23*	0.25
0.75	0.71	0.68	0.71	0.68	0.78	0.34	0.37	0.35	0.38	*0.33*	*0.33*	0.36
1.00	0.85	0.81	0.86	0.80	0.93	0.51	0.57	0.53	0.59	*0.48*	0.50	0.49
						P = 20						
b	VM	VM2	VMnc	VMnr	VMF	KM	KMnc	FM	FMnc	LGPSM	LGPSMnc	COVnc
0.00	0.21	0.19	0.22	0.18	0.23	0.22	0.22	0.22	0.22	0.20	0.20	*0.11*
0.25	0.39	0.37	0.39	0.36	0.41	0.24	0.25	0.25	0.26	0.24	0.23	*0.21*
0.50	0.56	0.53	0.57	0.52	0.62	0.35	0.38	0.37	0.40	0.35	*0.34*	0.35
0.75	0.77	0.73	0.79	0.64	0.84	0.65	0.67	0.66	0.71	0.57	0.56	*0.51*
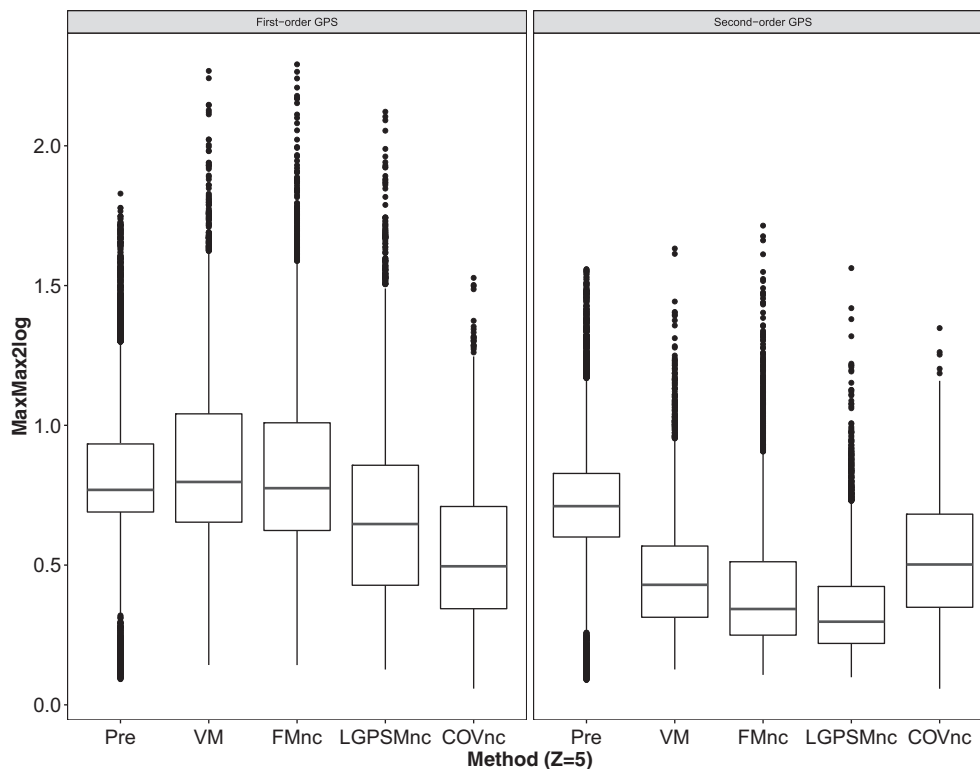
FIGURE 6 MaxMax2log for a prematched cohort across matching algorithms, $Z = 5$

Similar to $Z = 3$, initial covariate bias explains the highest portion of variability in MaxMax2SB, accounting for at least 25% of the variability for each algorithm (data not shown). The number of covariates, P, explains the second-largest portion of the variability in MaxMax2SB for most of the algorithms. The ratio of treatment-group sample sizes explains the third-largest portion of the variability in MaxMax2SB.

Table 5 shows the median MaxMax2SB for different levels of b, P, and the two GPS models. Overall trends for b, P, and the GPS model are similar to those observed for $Z = 3$. However, VM algorithms have the worse performance with increasing b. The medians MaxMax2SB for these algorithms are larger than 0.20 for $b > 0$ across all levels of P. Algorithms that match on the Mahalanobis distance of the logit GPS or the Mahalanobis distance of a subset of the logit GPS have medians MaxMax2SB that exceed 0.20 for $b = 1$ when $P = 5$ and for $b \geq 0.50$ when $P \in \{10, 20\}$.

The MaxMax2log for sets generated using each of the VM, FMnc, LGPSMnc, and COVnc is depicted in Figure 6. Including the square of the covariates in the GPS model results in smaller MaxMax2log for all algorithms. VM has the highest median MaxMax2log overall for both GPS models. LGPSMnc with the second-order GPS model has the lowest median MaxMax2log and 95% of the configurations with less than 0.69.

Table 6 shows MaxMax2log based on different levels of σ_2^2, σ_3^2, and GPS order. Similar to $Z = 3$, MaxMax2log increases as the absolute difference between σ_2^2 and σ_3^2 increases. All of the methods that match on the GPS have medians MaxMax2log that are smaller than 0.53 when using the second-order GPS model. COVnc has median MaxMax2log that reaches 0.64.

3.5 | Matching for $Z = 10$ treatments

We implemented VM, FMnc, LGPSMnc, and COVnc for $Z = 10$. We only included the noncaliper versions of the latter three algorithms because for $Z = 5$, we observed larger biases when these algorithms were implemented with calipers.

The MaxMax2SB for sets generated using each of the four algorithms with a first-order GPS model is depicted in Figure 7. The median MaxMax2SB for VM is significantly larger than the 0.20 cutoff with only 4% of configurations yielding MaxMax2SB lower than 0.20. The median MaxMax2SB is 0.25 and 0.27 for FMnc and LGPSMnc, respectively. Comparing to configurations with $Z = 3$ and $Z = 5$, MaxMax2SB is trending upward for LGPSMnc. These results are

TABLE 6 Median MaxMax2log across levels of σ_2^2, σ_3^2, and generalized propensity score (GPS) model order, $Z = 5$

	σ_2^2	σ_3^2	VM	FMnc	LGPSMnc	COVnc
First-order GPS	0.5	0.5	0.75	0.69	0.63	0.65
	0.5	1.0	0.75	0.71	0.61	0.65
	0.5	2.0	1.25	1.17	0.98	0.67
	1.0	0.5	0.76	0.72	0.59	0.61
	1.0	1.0	0.32	0.31	0.28	0.36
	1.0	2.0	0.77	0.76	0.62	0.41
	2.0	0.5	1.25	1.20	0.95	0.64
	2.0	1.0	0.77	0.75	0.62	0.42
	2.0	2.0	0.78	0.76	0.63	0.41
Second-order GPS	0.5	0.5	0.42	0.40	0.34	0.60
	0.5	1.0	0.44	0.37	0.31	0.61
	0.5	2.0	0.53	0.40	0.35	0.64
	1.0	0.5	0.40	0.36	0.31	0.58
	1.0	1.0	0.29	0.26	0.24	0.36
	1.0	2.0	0.44	0.29	0.26	0.41
	2.0	0.5	0.53	0.39	0.34	0.63
	2.0	1.0	0.46	0.30	0.26	0.42
	2.0	2.0	0.36	0.29	0.26	0.43
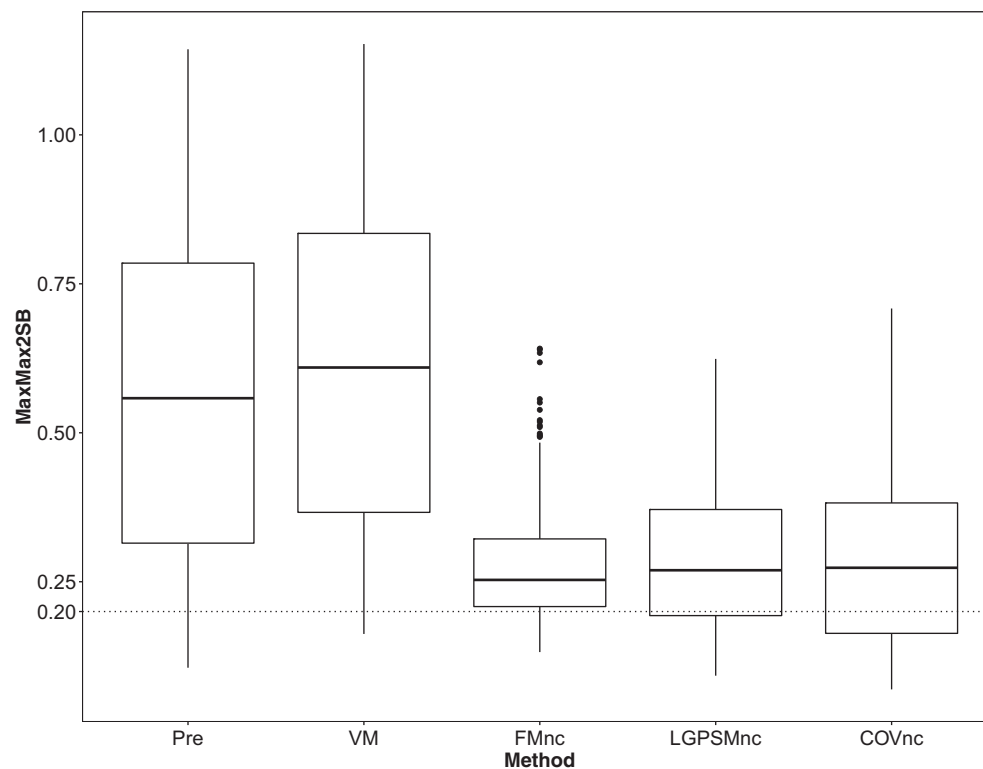
FIGURE 7 MaxMax2*SB* for a prematched cohort and by matching algorithms, $Z = 10$

similar to the results for matching on the Mahalanobis distance of the covariates for binary treatments, where the performance of matching on the Mahalanobis distance deteriorates with increasing P.[37,38] For FMnc, 19% of configurations yields MaxMax2*SB* below 0.20, with an interquartile range of (0.21, 0.32). VM is the only algorithm that used a caliper, and it yields a median *Prop.Matched* of 0.99. Similar to $Z = 3$ and $Z = 5$, MaxMax2*SB* was higher overall when using a second-order GPS model (data not shown).

TABLE 7 Median MaxMax2*SB* across levels of b, P, and generalized propensity score (GPS) model order, $Z = 10$ (smallest MaxMax2*SB* in italics)

First-Order GPS Model				
$P = 10$				
b	**VM**	**FMnc**	**LGPSMnc**	**COVnc**
0.00	0.22	0.19	0.14	*0.10*
0.25	0.39	0.20	0.16	*0.14*
0.50	0.62	0.23	0.23	*0.22*
0.75	0.84	*0.29*	0.31	0.31
1.00	1.03	*0.37*	0.42	0.42
$P = 20$				
b	**VM**	**FMnc**	**LGPSMnc**	**COVnc**
0.00	0.25	0.21	0.21	*0.19*
0.25	0.43	*0.24*	*0.24*	0.26
0.50	0.65	*0.30*	0.31	0.38
0.75	0.87	*0.42*	0.44	0.53
Second-Order GPS Model				
$P = 10$				
b	**VM**	**FMnc**	**LGPSMnc**	**COVnc**
0.00	0.23	0.20	0.18	*0.09*
0.25	0.37	0.21	0.21	*0.14*
0.50	0.60	0.24	0.29	*0.22*
0.75	0.80	*0.28*	0.37	0.31
1.00	0.99	*0.36*	0.47	0.42
$P = 20$				
b	**VM**	**FMnc**	**LGPSMnc**	**COVnc**
0.00	0.29	0.26	0.24	*0.18*
0.25	0.43	0.28	0.28	*0.25*
0.50	0.65	*0.33*	0.36	0.38
0.75	0.87	*0.48*	*0.48*	0.52

Table 7 shows the median MaxMax2*SB* for different levels of b, P, and the two GPS models. Similar to $Z = 5$, VM is more sensitive to an increase in b. For each algorithm, using either a first-order or a second-order GPS model does not appear to influence MaxMax2*SB* as much as it did for $Z \in \{3, 5\}$. FMnc and GPSnc have medians MaxMax2*SB* ≤ 0.20 only for $b = 0.00$ and $P = 10$, although the median MaxMax2*SB* exceeds 0.30 only for $b \geq 0.75$.

The MaxMax2log for sets generated using each of the VM, FMnc, LGPSMnc, and COVnc is depicted in Figure 8. Results are similar to those observed for $Z = 5$, where the medians MaxMax2log of VM, FMnc, and LGPSMnc are smaller when using a second-order GPS model, and the MaxMax2log for LGPSMnc is the smallest.

4 | AN APPLICATION TO THE EVALUATION OF NURSING HOME PERFORMANCE

4.1 | Nursing home data

The evaluation of providers′ performance based on patient outcomes has played an important role in the analysis and development of health care programs in the United States.[46] To estimate the effects of providers on patients′ outcomes, the populations of patients should be similar across all providers, also referred to as similar "case mix." While randomly assigning patients to different providers would be ideal, it would not be practical because patients may seek a provider that specializes in a particular condition.[47]

Rehospitalization rate is a common measure to compare the performance and quality of skilled nursing facilities (SNFs).[48,49] One possible method to compare SNF rehospitalization rate is risk adjustment using regression modeling.[47,50]
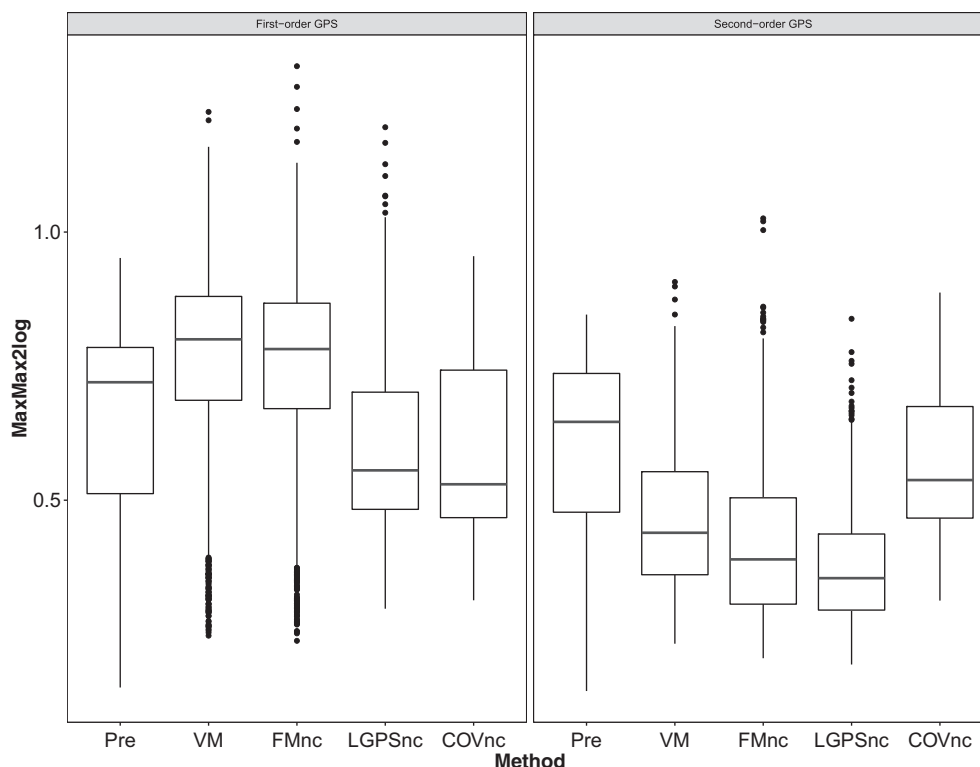
FIGURE 8 MaxMax2log for a prematched cohort across matching algorithms, $Z = 10$

However, regression-based risk adjustment is limited to cases for which the regression model is correctly specified or the covariates are balanced across the different providers.[35,51,52] Because the risk model is generally unknown and may suffer from model misspecification, it is important to ensure that the distribution of patients across providers is similar in order to avoid extrapolation and biased estimates.[46]

A different method to compare the rehospitalization rate among providers is to estimate each patient's outcome if they attended each of the available SNFs. To do this, one could match patients admitted to different SNFs who have similar sets of covariates and impute patients' missing rehospitalization status for an unadmitted SNF using the observed rehospitalization values of the matched units.

We apply the current and newly proposed matching algorithms to a data set consisting of Medicare enrollees linked to the Minimum Data Set who were discharged from Rhode Island Hospital over a 9-year period beginning on January 2, 1999, and assigned to a nearby nursing home in Rhode Island or Massachusetts. The outcome of interest is whether a patient was readmitted to Rhode Island Hospital within 30 days of their initial discharge. For each patient, demographic and clinical characteristics at discharge were recorded. The demographic characteristics include the patient's age, gender, and year of admission. The clinical characteristics include the primary ICD-9 diagnosis code for the hospital admission. We have also included in the analysis inflation-adjusted reimbursement for the hospital stay, which serves as a proxy for the intensiveness of the stay, as well as intensive care unit (ICU) use and SNF use in the 120 days prior to hospitalization. More than half of the 257 nursing homes in the data set had fewer than 10 patients. We focus on the five nursing homes with the most patients. The number of patients in these SNFs ranges from 670 to 1211, and we define the reference group as the nursing home with the largest number of patients (referred to as "nursing home 1").

We estimated the GPS vector using a multinomial logistic regression model and generated matched sets using five of the algorithms examined in previous sections: VM, FM, FMnc, LGPSMnc, and COVnc. Figure 9 reports the maximum absolute standardized pairwise bias, $Max2SB_p$, for each of the 30 baseline covariates in the original, unmatched sample and for the five matched samples generated by the aforementioned procedures. In the original sample, 12 of the 30 covariates have $Max2SB_p$ that exceeds 0.20, with the largest being prior nursing home use (0.77). LGPSMnc yields the most improved balance with 29 out of 30 covariates having $Max2SB_p$ smaller than 0.20 in the matched cohort, whereas FM and FMnc yield 27 out of 30 covariates having $Max2SB_p$ smaller than 0.20.
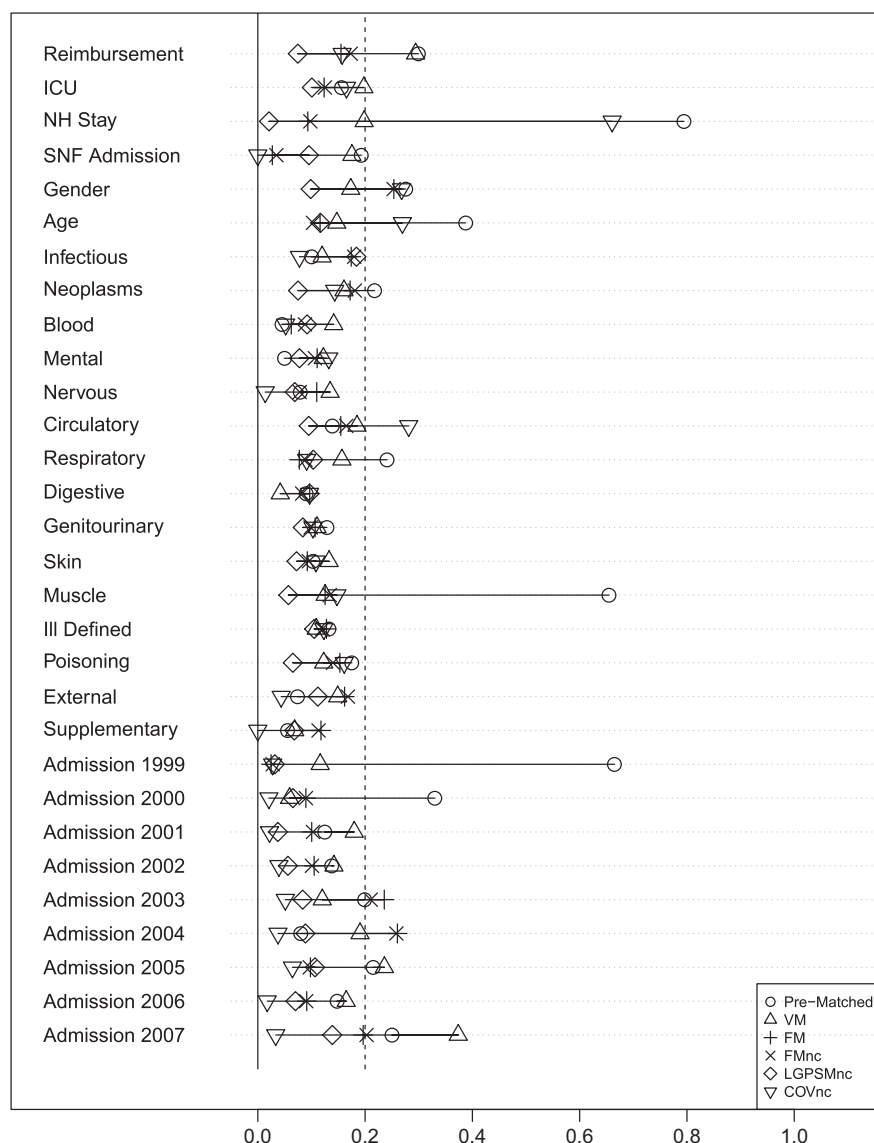
FIGURE 9 $Max2SB_p$ for baseline covariates across matching algorithms ($Z = 5$)

TABLE 8 MaxMax2SB and *Prop.Matched* for prematched and matched samples, nursing home data

Method	MaxMax2SB	*Prop.Matched*
Prematched	0.77	–
VM	0.30	0.99
FM	0.25	0.96
FMnc	0.24	1.00
LGPSMnc	0.23	1.00
COVnc	0.61	1.00

Table 8 shows MaxMax2SB and *Prop.Matched* across the five algorithms, compared to the prematched cohort. LGPSMnc yields the lowest MaxMax2SB of 0.23, followed by FMnc and FM with 0.24 and 0.25, respectively. VM and FM, the two algorithms that used a caliper, retain close to 100% of the reference group units in the matched cohort. Matching on the Mahalanobis distance of the covariates, COVnc, yields the largest MaxMax2SB.

Estimated differences in 30-day patient rehospitalization rates are given in Table 9 using each of the five matching algorithms for the five nursing homes. Pairwise comparisons between nursing homes 1 and 4 yielded uniformly positive

TABLE 9 Estimated outcomes (standard errors in parenthesis) for nursing home data, $Z = 5$

| Algorithm | Nursing Home Comparison | | | |
	1 vs 2	1 vs 3	1 vs 4	1 vs 5
VM	0.02 (0.02)	−0.02 (0.02)	0.04 (0.017)*	0.04 (0.02)*
FM	0.00 (0.02)	0.01 (0.02)	0.06 (0.01)*	0.04 (0.01)*
FMnc	0.00 (0.02)	0.01 (0.02)	0.06 (0.01)*	0.04 (0.01)*
LGPSMnc	0.01 (0.02)	0.06 (0.02)*	0.07 (0.02)*	0.02 (0.02)
COVnc	0.03 (0.02)	0.04 (0.02)*	0.07 (0.02)*	0.06 (0.02)*

*Significant at the 0.05 level.

and statistically significant differences in 30-day rehospitalization rates, with the effect size ranging from a 4.42% (VM) to a 7.28% (COVnc) higher rehospitalization rate for nursing home 1. Pairwise comparisons between nursing homes 1 and 2 yielded uniformly small and nonstatistically significant differences in 30-day rehospitalization rates.

Notable differences in estimated outcomes obtained from different matching algorithms appear when comparing nursing homes 1 and 3 and nursing homes 1 and 5. When comparing nursing homes 1 and 3, FM and FMnc yield smaller pairwise differences in 30-day rehospitalization rates in comparison to the other methods. When comparing nursing homes 1 and 5, LGPSMnc yields a smaller pairwise difference in 30-day rehospitalization rates, whereas VM, FM, and FMnc all yield an estimated difference of approximately 3.7%.

We further detail the comparisons of nursing home 1 versus nursing homes 2 through 5 by comparing standardized pairwise bias for each covariate. Tables A1-A4 in the Appendix compare SB_{p1k}, $k \in \{2, 3, 4, 5\}$, for each matching algorithm. In matching nursing home 1 patients to nursing home 3 patients, FM and FMnc yield greater SB_{p13} for the reimbursement and admission years 2003 and 2007, and COVnc yields greater SB_{p13} for gender, age, and the circulatory disease and poisoning indicators. In matching nursing home 1 patients to nursing home 5 patients, FM and FMnc yield greater SB_{p15} for the ICU and gender covariates, and COVnc yields greater SB_{p15} for the age, ICU, circulatory disease indicator, and poisoning indicator, in comparison to the other methods. LGPSMnc yields similar or smaller SB_{p13} and SB_{p15} for all covariates, findings that are similar to the simulation results for $Z = 5$, in which LGPSMnc displays smaller MaxMax2SB on average. With the exception of COVnc, SB_{p12} and SB_{p14} are less than 0.20 for all of the covariates and for all of the matching algorithms that were examined.

5 | DISCUSSION

Many applications in public health, medicine, and social sciences involve comparing multiple treatment groups. Matching is a useful tool for researchers looking to make causal statements between treatment groups. We have proposed several new matching algorithms when evaluating multiple treatments. We compared the performance of the newly proposed algorithms to previously proposed matching algorithms in terms of bias reduction in the covariates' distributions in different treatment groups.

VM is shown to result in the lowest covariate bias in comparison to other methods when estimating causal effects with three treatments, similar to the findings for three treatments in the work of Lopez and Gutman.[8] Using simulations, we examined several extensions of VM that use different clustering methods, distance measures, and calipers. When comparing three treatments, VM without replacement has the smallest bias; however, it also discards a large number of units receiving the reference treatment, thus sacrificing generalizability for minor gains in bias reductions. Moreover, as the number of treatment groups increases, VM and its extensions based on the linear GPS have the worst performance in terms of bias reduction compared to the other methods that were examined.

KM and FM are matching algorithms that match on the Mahalanobis distance of pairs of GPS vector components, within strata created by k-means or fuzzy clustering of the remaining components, respectively. While simulations show that FM had slightly better performance in terms of balancing covariates, the differences in performance were slight, especially when no caliper was used (KMnc and FMnc). The proportion of reference group units matched is generally higher when using FM, because fuzzy clustering allows for units to be assigned to multiple clusters. Each of these algorithms performs significantly better than VM-like algorithms for five or more treatments, indicating that matching on the Mahalanobis distance within clusters results in better covariate balance than using the linear GPS.

For three and five treatment groups, when there is a small number of covariates or the bias of the covariates is small, matching on the Mahalanobis distance of the covariates (COVnc) provides the best method in terms of bias reduction. However, the performance of this method deteriorates significantly as the number of covariates with initial biases increases. Thus, this method is not recommended for many practical applications, where a large number of covariates with different degrees of initial biases is expected. Matching on the Mahalanobis distance of the logit GPS vector with and without a caliper (LGPSM and LGPSMnc) generally has the largest reduction in bias while maintaining most of the units in the reference group. Matching with a caliper on the Mahalanobis distance of the logit GPS vector (LGPSM) has slightly lower average bias, but it discards more units than matching without a caliper. FM and FMnc have similar trends to LGPSM and LGPSMnc, but with slightly larger covariates' bias.

For 10 treatment groups, FMnc displays the largest reduction in covariates' bias. This implies that as the number of treatments increases, matching on the Mahalanobis distance of the logit GPS vector becomes less effective. Thus, our recommendation with more than five treatments is to rely on methods such as FM or FMnc that partition the cohort into clusters using fuzzy clustering and match units using the Mahalanobis distance within these clusters. In general, as the number of treatments increases, balancing the covariates across all treatment groups is a harder task, and imbalances across treatment groups may still remain even after matching. A possible limitation with matching for 10 or more treatment groups is that because of the large number of comparisons being made, it can be difficult to detect significant differences in treatment outcomes after adjustments for multiple comparisons. This problem arises for both randomized and nonrandomized studies, and it depends on the overall effect sizes, variability of the outcomes in the populations, and the number of units within the different treatment groups.

When matching on a GPS model that includes squared covariates, larger biases in the original covariates are observed, in comparison to a GPS model that only included the covariates. This is because the GPS model with squared terms balances more variables. However, when using a GPS model that includes squared terms, the variance ratios between different treatment groups are significantly lower. For binary treatment, differences in covariates' variance are shown to have a large influence on the operating characteristics of causal estimates.[52] Thus, it is important to ensure that the ratios of variances are similar across treatment groups. Our simulations show that this can be accomplished by including squared terms in the GPS model with relatively minor increases in covariates' biases.

Applying the different matching algorithms to examine the effect of admission to one of five SNFs reinforced the simulation results. Matching on the Mahalanobis distance of the logit GPS vector without a caliper has the largest reduction in bias while maintaining all of the units in the largest SNF. On the basis of these results, we can identify that residing in certain SNFs can result in a lower chance of being readmitted to a hospital within 30 days. One limitation of this analysis is that it assumes that all of the covariates that influence admission to one of the SNFs are available or, formally, that the assignment mechanism is unconfounded. This is a limitation of any observational study, and it will generally have a larger effect on many of the currently employed risk adjustment techniques as well. Developing sensitivity analyses to this assumption with multiple treatment groups is an area of future research. In addition, health provider profiling generally involves the comparison of thousands of SNFs simultaneously, and development of matching methods in this scenario is also an area of further research. Nonetheless, most patients can only choose from a limited number of possible SNFs. Thus, the proposed matching procedures can provide a viable analysis option for estimating the causal effects of residing in one SNF versus others.

In conclusion, we compared several new and previously proposed matching procedures when attempting to estimate the causal effects from observational studies with multiple treatments. On the basis of our simulations, for three to five treatments, matching on the Mahalanobis distance of the logit GPS vector with (LGPSM) or without a caliper (LGPSMnc) provides the largest reduction in initial covariate bias. However, as the number of treatments increases, matching methods that are based on the Mahalanobis distance and fuzzy clustering (FM and FMnc) provide better reduction in initial covariates' bias. In addition, matching methods that include the squared term of the covariates in the GPS model result in greater reduction in covariates' variance ratios between treatment groups at the expense of slightly larger covariates' bias.

not necessarily represent the views of the Patient-Centered Outcomes Research Institute, its Board of Governors or Methodology Committee.

FINANCIAL DISCLOSURE

None reported.

CONFLICT OF INTEREST

The authors declare no potential conflict of interests.

SUPPORTING INFORMATION

Additional Supporting Information may be found online in the supporting information tab for this article. The online Supplementary Material includes R code used to run each of the matching algorithms that are examined in this manuscript. Data are available in a synthetic format at `https://github.com/ScotinaStats`.

ORCID

Anthony D. Scotina 🔟 https://orcid.org/0000-0003-1301-3209

REFERENCES

1. Barthel FMS, Parmar MKB, Royston P. How do multi-state, multi--arm parallel group design – a reanalysis of 4 trials. *Trials*. 2009;10:1-10.
2. Parmar MKB, Carpender J, Sydes MR. More multiarm randomised trials of superiority are needed. *Lancet*. 2014;384:283-284.
3. Ghosh P, Liu L, Senchaudhuri P, Gao P, Mehta C. Design and monitoring of multi-arm multi-stage clinical trials. *Biometrics*. 2017;73(4):1289-1299.
4. Jaki T. Multi-arm clinical trials with treatment selection: what can be gained and at what price? *Clin Investig*. 2015;5(4):393-399.
5. Wason JMS, Stecher L, Mander AP. Correcting for multiple-testing in multi-arm trials: is it necessary and is it done? *Trials*. 2014;15:1-7.
6. Cook RJ, Farewell VT. Multiplicity concerns in the design and analysis of clinical trials. *J Royal Stat Soc Ser A*. 1996;159:93-110.
7. Bender R, Lange S. Adjusting for multiple testing—when and how? *Clin Epidemiol*. 2001;54:343-349.
8. Lopez MJ, Gutman R. Estimation of causal effects with multiple treatments: a review and new ideas. *Stat Sci*. 2017;32:432-454.
9. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika*. 1983;70:41-55.
10. Rubin DB, Thomas N. Characterizing the effect of matching using linear propensity score methods with normal distributions. *Biometrika*. 1992;79:797-809.
11. Rubin DB, Stuart EA. Affinely invariant matching methods with discriminant mixtures of proportional ellipsoidally symmetric distributions. *Ann Stat*. 2006;34:1814-1826.
12. Dehejia RH, Wahba S. Propensity score matching methods for nonexperimental causal studies. *Rev Econ Stat*. 2002;84:151-161.
13. Monahan KC, Lee JM, Steinberg L. Revisiting the impact of part-time work on adolescent adjustment: distinguishing between selection and socialization using propensity score matching. *Child Dev*. 2010;82:96-112.
14. Resa M, Zubizarreta JR. Evaluation of subset matching methods and forms of covariate balance. *Statist Med*. 2016;35:4961-4979.
15. Rosenbaum PR. Optimal matching for observational studies. *J Am Stat Assoc*. 1989;84(408):1024-1032.
16. Zubizarreta JR. Using mixed integer programming for matching in an observational study of kidney failure after surgery. *J Am Stat Assoc*. 2012;107:1360-1371.
17. Imbens GW. The role of the propensity score in estimating dose-response functions. *Biometrika*. 2000;87:706-710.
18. Imai K, Dyk DA. Causal inference with general treatment regimes: generalizing the propensity score. *J Am Stat Assoc*. 2004;99(467):854-866.
19. Cochran WG, Rubin DB. Controlling bias in observational studies: a review. *Sankhya Ser A*. 1973;35:417-446.
20. Rubin DB. Matching to remove bias in observational studies. *Biometrics*. 1973;29:159-183.
21. Rubin DB. The use of matched sampling and regression adjustment to remove bias in observational studies. *Biometrics*. 1973;29:185-203.
22. Rosenbaum PR. *Observational Studies*. 2nd ed. New York, NY: Springer; 2002.
23. Hansen BB. Full matching in an observational study of coaching for the SAT. *J Am Stat Assoc*. 2004;99:609-618.
24. Karp RM, Miller RE, Thatcher JW. Reducibility among combinatorial problems. In: Miller RE, Thatcher JW. *Complexity of Computer Computations*. Boston, MA: Springer; 1972:85-103.

25. Rubin DB. Estimating causal effects of treatments in randomized and nonrandomized studies. *J Educ Psychol*. 1974;56(5):688-701.

26. Rubin DB. Bayesian inference for causal effects: the role of randomization. *Ann Stat*. 1978;6(1):34-58.

27. Holland PW. Statistics and causal inference. *J Am Stat Assoc*. 1986;81(396):945-960.

28. McCaffrey DF, Griffin BA, Almirall D, Slaughter ME, Ramchand R, Burgette LF. A tutorial on propensity score estimation for multiple treatments using generalized boosted models. *Statist Med*. 2013;32(19):3388-3414.

29. Imbens GW, Rubin DB. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. New York, NY: Cambridge University Press; 2015.

30. Yang S, Imbens GW, Cui Z, Faries DE, Kadziola Z. Propensity score matching and subclassification in observational studies with multi-level treatment. *Biometrics*. 2016;72:1055-1065.

31. Rosenbaum PR, Rubin DB. Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *Am Stat*. 1985;39:33-38.

32. Rubin DB. Using propensity scores to help design observational studies: application to the tobacco litigation. *Health Serv Outcomes Res Methodol*. 2001;2(3):169-188.

33. Krzanowski WJ. *Principles of Multivariate Analysis: A User's Perspective*. New York, NY: Oxford University Press; 2000.

34. Steiner PM, Cook DL. Matching and propensity scores. In: *The Oxford Handbook of Quantitative Methods, Volume 1: Foundations*. New York, NY: Oxford University Press; 2013.

35. Rubin DB. Using multivariate matched sampling and regression adjustment to control bias in observational studies. *J Am Stat Assoc*. 1979;74(366):318-328.

36. Rubin DB. Randomization analysis of experimental data: the Fisher randomization test comment. *J Am Stat Assoc*. 1980;75:591-593.

37. Gu X, Rosenbaum PR. Comparison of multivariate matching methods: structures, distances, and algorithms. *J Comput Graph Stat*. 1993;2:405-420.

38. Stuart EA. Matching methods for causal inference: a review and a look forward. *Stat Sci*. 2010;25(1):1-21.

39. Cohen J. *Statistical Power Analysis for the Behavioral Sciences*. 2nd ed. Abingdon, UK: Routledge; 1988.

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

41. Bezdek JC, Ehrlich R, Full W. The fuzzy c-means clustering algorithm. *Comput Geosci*. 1984;10:191-203.

42. R Studio Team. RStudio: Integrated Development Environment for R RStudio Inc., Boston, MA. 2015.

43. Sekhon J. Multivariate and propensity score matching software with automated balance optimization: the matching package for R. *J Stat Softw*. 2011;42(7):1-52.

44. Rubin DB. The design versus the analysis of observational studies for causal effects: parallels with the design of randomized trials. *Statist Med*. 2007;26:20-36.

45. Cangul MZ, Chretien YR, Gutman R, Rubin DB. Testing treatment effects in unconfounded studies under model misspecification: logistic regression, discretization, and their combination. *Statist Med*. 2009;28(20):2531-2551.

46. Ash AS, Fienberg SF, Louis TA, Normand ST, Sukel TA, Utts J. Statistical issues in assessing hospital performance. https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/HospitalQualityInits/Downloads/Statistical-Issues-in-Assessing-Hospital-Performance.pdf. Accessed November 2, 2012.

47. Huang I, Frangakis C, Dominici F, Diette GB, Wu AW. Application of a propensity score approach for risk adjustment in profiling multiple physician groups on asthma care. *Health Serv Res*. 2005;40:253-278.

48. Ouslander JG, Lamb G, Perloe M, et al. Potentially avoidable hospitalizations of nursing home residents: frequency, causes, and costs. *J Am Geriatr Soc*. 2010;58:627-635.

49. Ouslander JG, Berenson RA. Reducing unnecessary hospitalizations of nursing home residents. *N Engl J Med*. 2011;365:1165-1167.

50. Neuman MD, Wirtalla C, Werner RM. Association between skilled nursing facility quality indicators and hospital readmissions. *J Am Med Assoc*. 2014;312:1542-1551.

51. Dehejia RH, Wahba S. Causal effects in nonexperimental studies: reevaluating the evaluation of training programs. *J Am Stat Assoc*. 1999;94:1053-1062.

52. Gutman R, Gutman H. Monitoring methods in the surgical arena: assumptions and potential pitfalls. *Eur J Surg Oncol*. 2013;39:575-583.

APPENDIX

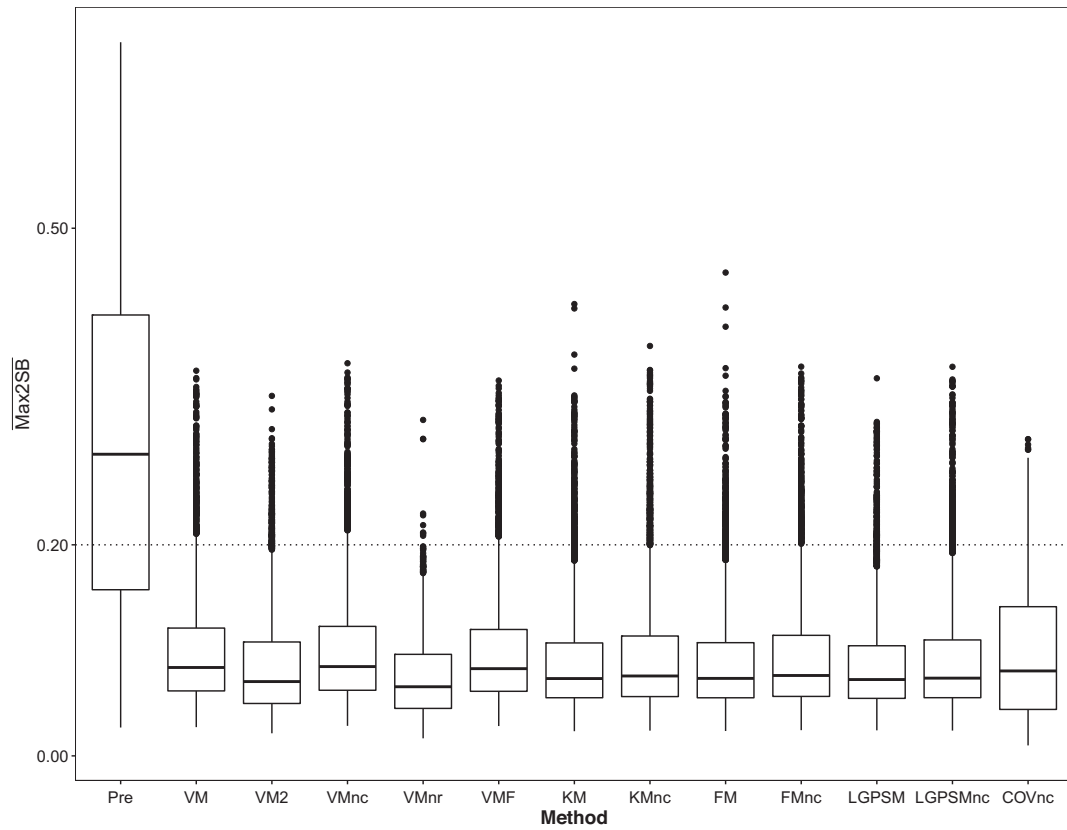## A ADDITIONAL TABLES AND FIGURES

FIGURE A1 $\overline{\text{Max2}SB}$ for a prematched cohort and by matching algorithms, $Z = 3$
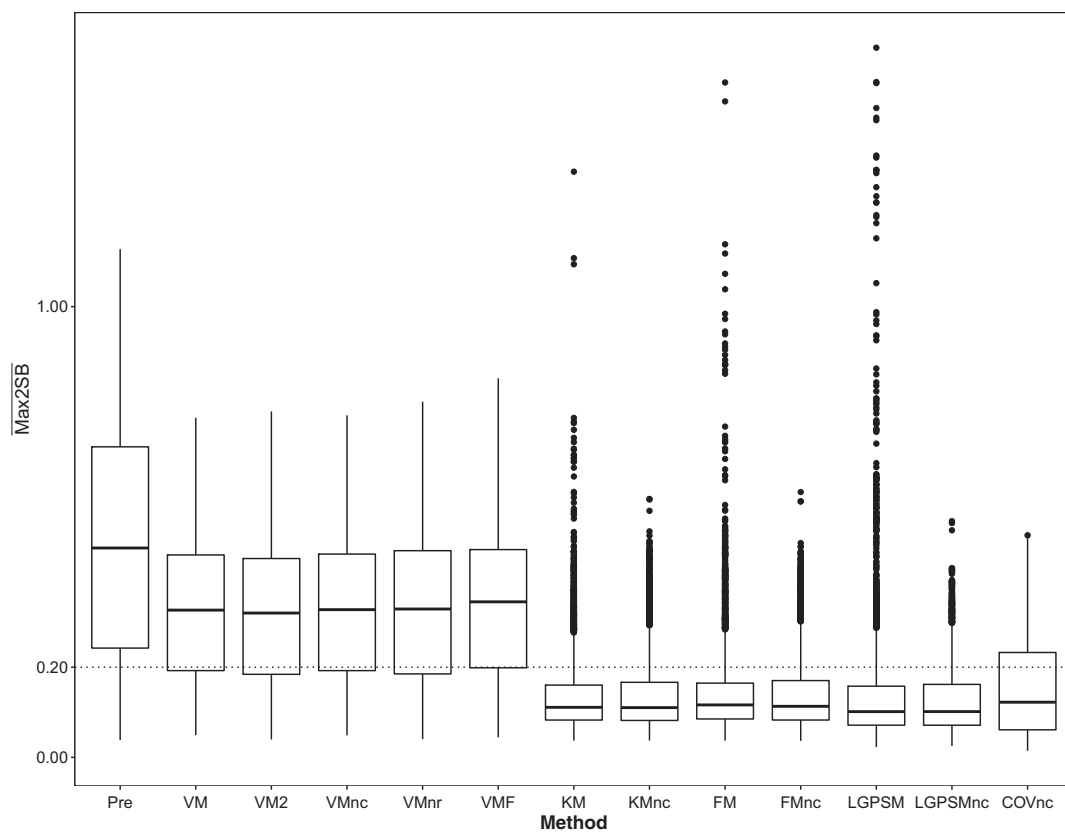
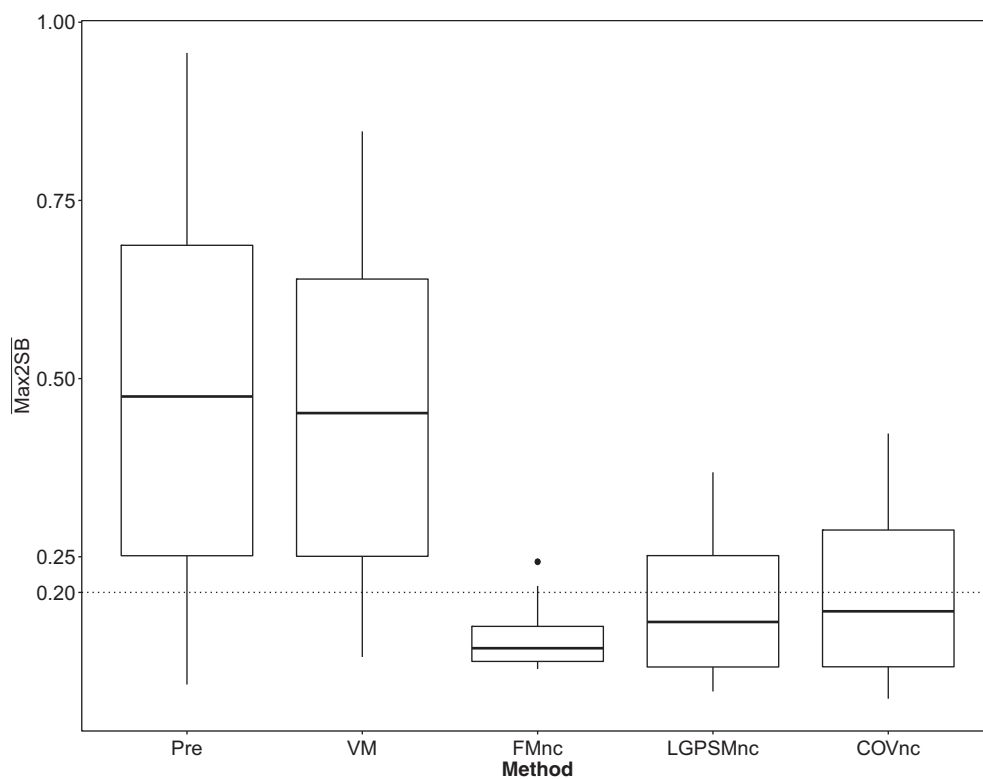**FIGURE A2** $\overline{\text{Max2SB}}$ for a prematched cohort and by matching algorithms, $Z = 5$

FIGURE A3 $\overline{\text{Max2SB}}$ for a prematched cohort and by matching algorithms, $Z = 10$

TABLE A1 SB_{p12} for nursing home data

Variable	VM	FM	FMnc	LGPSMnc	COVnc
Reimbursement	0.07	0.01	0.01	0.05	0.16
ICU	0.02	0.00	0.02	0.05	0.13
NH Stay	0.02	0.01	0.01	0.03	0.07
SNF Admission	0.10	0.03	0.03	0.00	0.00
Gender	0.05	0.01	0.01	0.02	0.23
Age	0.13	0.01	0.00	0.08	0.21
Infectious	0.00	0.05	0.06	0.03	0.00
Neoplasms	0.00	0.07	0.07	0.01	0.02
Blood	0.06	0.02	0.03	0.06	0.02
Mental	0.02	0.01	0.02	0.00	0.09
Nervous	0.00	0.01	0.02	0.05	0.01
Circulatory	0.03	0.02	0.01	0.04	0.20
Respiratory	0.01	0.05	0.04	0.07	0.03
Digestive	0.00	0.00	0.01	0.07	0.04
Genitourinary	0.02	0.01	0.02	0.02	0.03
Skin	0.05	0.06	0.06	0.03	0.05
Muscle	0.01	0.02	0.01	0.03	0.04
Ill Defined	0.08	0.03	0.03	0.00	0.12
Poisoning	0.02	0.01	0.00	0.01	0.09
External	0.06	0.01	0.02	0.00	0.00
Supplementary	0.03	0.02	0.04	0.05	0.00
Admission 1999	0.04	0.02	0.01	0.02	0.00
Admission 2000	0.04	0.01	0.01	0.03	0.00
Admission 2001	0.02	0.05	0.05	0.03	0.01
Admission 2002	0.03	0.02	0.02	0.02	0.02
Admission 2003	0.03	0.02	0.01	0.07	0.01
Admission 2004	0.01	0.03	0.03	0.02	0.02
Admission 2005	0.08	0.03	0.03	0.02	0.03
Admission 2006	0.07	0.03	0.04	0.10	0.02
Admission 2007	0.03	0.00	0.01	0.00	0.01

TABLE A2 SB_{p13} for nursing home data

Variable	VM	FM	FMnc	LGPSMnc	COVnc
Reimbursement	0.07	0.18	0.12	0.04	0.13
ICU	0.02	0.10	0.08	0.10	0.07
NH Stay	0.00	0.06	0.06	0.03	0.23
SNF Admission	0.05	0.01	0.01	0.03	0.00
Gender	0.04	0.02	0.01	0.06	0.12
Age	0.01	0.02	0.02	0.08	0.28
Infectious	0.02	0.05	0.04	0.06	0.07
Neoplasms	0.06	0.06	0.04	0.02	0.04
Blood	0.05	0.00	0.01	0.01	0.01
Mental	0.02	0.03	0.03	0.00	0.08
Nervous	0.05	0.02	0.05	0.13	0.01
Circulatory	0.04	0.01	0.02	0.01	0.11
Respiratory	0.07	0.02	0.01	0.09	0.06
Digestive	0.03	0.06	0.05	0.06	0.09
Genitourinary	0.06	0.02	0.02	0.05	0.01
Skin	0.06	0.07	0.05	0.00	0.08
Muscle	0.14	0.11	0.09	0.05	0.06
Ill Defined	0.07	0.01	0.02	0.06	0.11
Poisoning	0.04	0.01	0.01	0.03	0.15
External	0.04	0.07	0.08	0.04	0.06
Supplementary	0.19	0.13	0.11	0.03	0.00
Admission 1999	0.02	0.05	0.04	0.01	0.00
Admission 2000	0.01	0.02	0.02	0.02	0.00
Admission 2001	0.01	0.03	0.04	0.02	0.02
Admission 2002	0.01	0.09	0.07	0.01	0.01
Admission 2003	0.11	0.05	0.04	0.16	0.00
Admission 2004	0.08	0.18	0.17	0.08	0.01
Admission 2005	0.17	0.12	0.11	0.14	0.01
Admission 2006	0.09	0.07	0.05	0.06	0.01
Admission 2007	0.12	0.14	0.13	0.12	0.02

TABLE A3 SB_{p14} for nursing home data

Variable	VM	FM	FMnc	LGPSMnc	COVnc
Reimbursement	0.16	0.09	0.03	0.01	0.00
ICU	0.14	0.00	0.03	0.04	0.16
NH Stay	0.26	0.01	0.01	0.01	0.38
SNF Admission	0.03	0.02	0.06	0.03	0.00
Gender	0.04	0.00	0.00	0.07	0.09
Age	0.13	0.01	0.03	0.05	0.06
Infectious	0.05	0.02	0.01	0.00	0.05
Neoplasms	0.04	0.05	0.05	0.01	0.04
Blood	0.04	0.04	0.03	0.07	0.03
Mental	0.09	0.04	0.04	0.03	0.13
Nervous	0.01	0.03	0.03	0.00	0.00
Circulatory	0.12	0.03	0.03	0.13	0.29
Respiratory	0.05	0.02	0.02	0.01	0.10
Digestive	0.04	0.01	0.00	0.09	0.08
Genitourinary	0.01	0.07	0.07	0.03	0.10
Skin	0.08	0.08	0.07	0.11	0.11
Muscle	0.00	0.06	0.03	0.03	0.08
Ill Defined	0.03	0.08	0.08	0.04	0.12
Poisoning	0.02	0.02	0.02	0.05	0.06
External	0.03	0.03	0.02	0.02	0.04
Supplementary	0.01	0.06	0.05	0.00	0.00
Admission 1999	0.06	0.02	0.02	0.06	0.01
Admission 2000	0.06	0.00	0.01	0.04	0.02
Admission 2001	0.16	0.05	0.05	0.05	0.00
Admission 2002	0.01	0.03	0.02	0.05	0.00
Admission 2003	0.05	0.05	0.03	0.07	0.02
Admission 2004	0.03	0.02	0.02	0.09	0.00
Admission 2005	0.10	0.12	0.11	0.06	0.00
Admission 2006	0.04	0.03	0.04	0.03	0.01
Admission 2007	0.05	0.10	0.10	0.05	0.01

TABLE A4 SB_{p15} for nursing home data

Variable	VM	FM	FMnc	LGPSMnc	COVnc
Reimbursement	0.07	0.01	0.07	0.04	0.12
ICU	0.07	0.10	0.12	0.05	0.16
NH Stay	0.05	0.06	0.06	0.01	0.02
SNF Admission	0.04	0.01	0.00	0.03	0.00
Gender	0.04	0.13	0.12	0.01	0.04
Age	0.05	0.07	0.04	0.04	0.16
Infectious	0.03	0.01	0.01	0.13	0.05
Neoplasms	0.07	0.07	0.07	0.08	0.10
Blood	0.04	0.05	0.06	0.03	0.05
Mental	0.01	0.02	0.01	0.02	0.08
Nervous	0.01	0.10	0.10	0.06	0.02
Circulatory	0.07	0.03	0.03	0.07	0.22
Respiratory	0.02	0.03	0.03	0.01	0.05
Digestive	0.05	0.07	0.08	0.03	0.07
Genitourinary	0.05	0.02	0.02	0.03	0.00
Skin	0.05	0.04	0.05	0.01	0.10
Muscle	0.02	0.02	0.04	0.01	0.07
Ill Defined	0.02	0.09	0.08	0.05	0.09
Poisoning	0.04	0.08	0.06	0.04	0.17
External	0.00	0.02	0.03	0.02	0.01
Supplementary	0.02	0.03	0.01	0.03	0.00
Admission 1999	0.03	0.01	0.01	0.02	0.01
Admission 2000	0.01	0.02	0.02	0.01	0.01
Admission 2001	0.04	0.03	0.04	0.05	0.01
Admission 2002	0.05	0.06	0.05	0.02	0.02
Admission 2003	0.07	0.13	0.11	0.05	0.02
Admission 2004	0.10	0.07	0.07	0.02	0.04
Admission 2005	0.14	0.13	0.13	0.00	0.03
Admission 2006	0.02	0.05	0.05	0.02	0.02
Admission 2007	0.16	0.06	0.07	0.01	0.01

B NURSING HOME DATA CODEBOOK

- Reimbursement—inflation-adjusted reimbursement (US $2000)
- ICU—any ICU during hospitalization
- NH Stay—nursing home stay within 120 days before inpatient admission
- SNF Admission—SNF admission (yes/no)
- Gender—beneficiary sex (1=male, 2=female)
- Age—beneficiary age
- Infectious—primary ICD-9 diagnosis 001–139: infectious and parasitic diseases
- Neoplasms—primary ICD-9 diagnosis 140–239: neoplasms
- Blood—primary ICD-9 diagnosis 280–289: diseases of the blood and blood-forming organs
- Mental—primary ICD-9 diagnosis 290–319: mental disorders
- Nervous—primary ICD-9 diagnosis 320–389: diseases of the nervous system and sense organs
- Circulatory—primary ICD-9 diagnosis 390–459: diseases of the circulatory system
- Respiratory—primary ICD-9 diagnosis 460–519: diseases of the respiratory system
- Digestive—primary ICD-9 diagnosis 520–579: diseases of the digestive system
- Genitourinary—primary ICD-9 diagnosis 580–629: diseases of the genitourinary system
- Skin—primary ICD-9 diagnosis 680–709: diseases of the skin and subcutaneous tissue
- Muscle—primary ICD-9 diagnosis 710–739: diseases of the musculoskeletal system and connective tissue
- Ill Defined—primary ICD-9 diagnosis 780–799: symptoms, signs, and ill-defined conditions
- Poisoning—primary ICD-9 diagnosis 800–999: injury and poisoning
- External—primary ICD-9 diagnosis E: external causes of injury
- Supplementary—primary ICD-9 diagnosis V: supplemental classification
- Admission 1999—nursing home year of admission in 1999
- Admission 2000—nursing home year of admission in 2000
- Admission 2001—nursing home year of admission in 2001
- Admission 2002—nursing home year of admission in 2002
- Admission 2003—nursing home year of admission in 2003
- Admission 2004—nursing home year of admission in 2004
- Admission 2005—nursing home year of admission in 2005
- Admission 2006—nursing home year of admission in 2006
- Admission 2007—nursing home year of admission in 2007