Research Article

Statistics in Medicine

Received XXXX

(www.interscience.wiley.com) DOI: 10.1002/sim.0000

Evaluation of Subset Matching Methods and Forms of Covariate Balance

María de los Angeles Resa[*] and José R. Zubizarreta[†]

This paper conducts a Monte Carlo simulation study to evaluate the performance of multivariate matching methods that select a subset of treatment and control observations. The matching methods studied are the widely used nearest neighbor matching with propensity score calipers, and the more recently proposed methods, optimal matching of an optimally chosen subset and optimal cardinality matching. The main findings are: (i) covariate balance, as measured by differences in means, variance ratios, Kolmogorov-Smirnov distances, and cross-match test statistics, is better with cardinality matching since by construction it satisfies balance requirements; (ii) for given levels of covariate balance, the matched samples are larger with cardinality matching than with the other methods; (iii) in terms of covariate distances, optimal subset matching performs best; (iv) treatment effect estimates from cardinality matching have lower RMSEs, provided strong requirements for balance, specifically, fine balance, or strength-k balance, plus close mean balance. In standard practice, a matched sample is considered to be balanced if the absolute differences in means of the covariates across treatment groups are smaller than 0.1 standard deviations. However, the simulation results suggest that stronger forms of balance should be pursued in order to remove systematic biases due to observed covariates when a difference in means treatment effect estimator is used. In particular, if the true outcome model is additive then marginal distributions should be balanced, and if the true outcome model is additive with interactions then low-dimensional joints should be balanced.

Keywords: Matched sampling; Observational studies; Propensity scores.

1. Introduction

In observational studies, matching methods are often used to approximate the ideal randomized experiment that would have been conducted if controlled experimentation had been feasible [1]. In these settings, the ultimate goal of matching is to free the comparisons of the outcomes in the treatment and control groups from biases due to differences in their observed covariates [2]. To achieve this goal, matching methods find samples of treated and control units with similar

[*] Ph.D. Student, Statistics Department, Columbia University, 1255 Amsterdam Avenue, 901 SSW, New York, NY 10027, Email: mdr2146@columbia.edu.
[†] Assistant Professor, Division of Decision, Risk and Operations, and Statistics Department, Columbia University, 3022 Broadway, 417 Uris Hall, New York, NY 10027-6902, Email: zubizarreta@columbia.edu.

The authors thank Winston Lin for useful comments and suggestions, and the support of a grant from the Alfred P. Sloan Foundation.

John Wiley & Sons, Ltd, The Atrium, Southern Gate, Chichester, West Sussex, PO19 8SQ, UK
[*] *Correspondence to: Journals Production Department, John Wiley & Sons, Ltd, The Atrium, Southern Gate, Chichester, West Sussex, PO19 8SQ, UK.*

or balanced observed covariate distributions [3]. Ideally, these matched samples will include all the available treated and control observations. However, if the distributions of the observed covariates are substantially different between treated and control samples before matching, as it tends to happen in observational studies, then including all the observations will result in poor covariate balance.

To address this problem, matching methods typically rely on the propensity score to remove observations that do not have an appropriate counterpart on the other treatment group.[§] One way to do this is to trim the sample before matching, discarding the observations outside the overlap region between the propensity score ranges of the two groups (see, e.g., [7]). Another widely used approach is to restrict the possible matches to be within a propensity score caliper. While these methods provide a broad solution to the problem, they are additions to methods for which discarding observations was not incorporated in the design of the problem.

Two recently proposed methods incorporate the selection of observations into their design and offer optimal solutions in some sense. The first method is optimal matching of an optimally chosen subset, or optimal subset matching for short [8]. This method pursues two specific goals: to minimize the total sum of covariate distances between the matched pairs, and to match as many pairs of treated and control units as possible. Often these two goals are at odds with each other, and the trade-off between them is regulated by means of a tuning parameter discussed and interpreted by [8].

The second method is optimal cardinality matching or, simply, cardinality matching [9]. This method finds the largest matched sample of treated and control units for which the covariate distributions are balanced as required by the researcher. With this method different forms of covariate balance can be achieved by design, including balance of means and higher order moments [10], balance of marginal distributions or fine balance [11], and balance of low dimensional joints or strength-k balance [12].

While there have been a number of simulation studies comparing different matching methods, most of these studies are centered on methods that rely on the propensity score, and to our knowledge none of them have considered these two optimal matching methods. For this reason perhaps is that these methods are not used much for research in medicine and other related disciplines, where they can play an important role because of their optimality guarantees in terms of covariate distances and balance. Furthermore, while there have been studies evaluating algorithms, calipers, distances, and structures in matching, few have placed emphasis on how covariates should be balanced. In practice, a widely used rule of thumb is to consider a matched sample to be balanced if the absolute standardized differences in means of the covariates across treatment groups are smaller than 0.1. However, we are not aware of a systematic study of this rule. Also, there are reasons to think that the balance criteria should be data- and estimator-specific. In particular, we are interested in studying questions like, in which cases would balancing only the means of the covariates be sufficient to remove biases when estimating the treatment effect with a simple difference in means? When should stronger forms of covariate balance, such as balance of marginal or low-dimensional joint distributions, be pursued? Although we believe these questions should ultimately be addressed with formal statistical theory, in this paper we conduct a Monte Carlo simulation study to provide initial answers. For this, in Section 2 we describe with more detail the matching methods that will be compared. In Section 3, we explain the specifics of the simulation study, and in Section 4 we show and discuss the simulation results. In Section 5 we explore additional considerations regarding the results. Finally, in Section 6 we conclude with a summary and remarks.

[§] As argued in [4], dropping units results in a smaller sample size and, it may appear, in a larger variance of the estimator, however better covariate balance may actually improve the efficiency of the estimator (see section 18.2 of [5]). Also, as argued in [6], dropping units may help to increase unit homogeneity which in turn can reduce sensitivity to biases due to unobserved covariates.

2. Matching methods

2.1. Greedy matching

Propensity score matching [13] is, in all likelihood, the most widely used matching method in medicine and related sciences. Accordingly, it has been studied extensively in the past (e.g., [14, 15, 16, 17, 18]), and thus, we considered it in our simulation study as a benchmark. A commonly used algorithm for propensity score matching is greedy or first-best nearest neighbor matching [19]. In its most basic form, this algorithm first sorts the treated units in terms of the estimated propensity score (from highest to lowest, lowest to highest, or randomly), and then matches the first treated unit to the closest available control, making it no longer available for matching for the rest of the treated units. Closeness here may be defined as the propensity score distance, which is the absolute difference of the estimated propensity scores for two units [20]. To avoid poor matches, a caliper may be added to the propensity score distance so that a control unit is matched to a treated unit only if it is within the caliper, and treated units for which there are no controls available within the caliper are discarded. In this manner, only a subset of the treated units are matched to controls. Simulation studies by [16, 17, 18] suggest that the best way to implement greedy nearest neighbor matching is by matching the treated units in random order and without replacement, using a linear propensity score distance, and imposing a caliper of 0.2 times the standard deviation of the linear propensity score. We followed these recommendations in our implementation of this method.

2.2. Optimal subset matching

Greedy matching does not, in general, minimize the total sum of distances between matched units (see chapter 10 of [21] for an example). On the contrary, optimal matching [22] finds the assignment of treated and control units that minimizes this global distance. In observational studies, the optimal matching problem can be cast as an assignment problem [23], a special case of the minimum cost flow problem [24], that in turn can be written as a linear program [25]. While it is possible to solve the assignment problem using the simplex algorithm, there exist specific algorithms such as the Hungarian algorithm [26] or the auction algorithm [27] that can better exploit the assignment problem's particular structure. This is important because these algorithms can be solved "quickly," or more formally, in polynomial time; that is, in a number of arithmetic operations that is characterized by a polynomial function of certain parameters of the problem (as opposed, say, to an exponential function [28]).

Typically, optimal matching uses all the available treated units and does not have the flexibility to discard some treated units in the case where they are very hard to match. A recently proposed matching method that selects a subset of units is optimal subset matching [8]. This is an elegant solution to the optimal subset matching problem, which consists of formulating the problem as an assignment problem on a modified matrix of distances between treated and control units; see [8] for details. As mentioned in the introduction, this method pursues two specific goals: to minimize the total sum of covariate distances between the matched units, and to match as many pairs of treated and control units as possible. Often these two goals are at odds with each other, but this trade-off is regulated by means of a prespecified covariate distance threshold $\tilde{\delta}$. For a given $\tilde{\delta}$, this method "prefers more treated subjects if their average increase in distance is less than $\tilde{\delta}$ and prefers fewer treated subjects if their average increase in distance is more than $\tilde{\delta}$, so $\tilde{\delta}$ is the distance at which there is indifference" [8].[¶]

In our simulation study we used optimal subset matching with the prespecified covariate distance threshold $\tilde{\delta}$ set up at the 20% quantile of the distance matrix before adding the caliper. For comparability with greedy matching we used the linear propensity score distance. We implemented this method using the R functions in the supplementary materials of [8], which in turn uses the R function pairmatch in the optmatch package [29].

[¶] This is achieved by solving the assignment problem with an augmented distance matrix, in which a certain number of columns, all with the value $\tilde{\delta}$, are added to the original treated-control distance matrix. These columns can be thought of as additional controls for which the distance to every treated unit is $\tilde{\delta}$, and the pairs that involve any of these controls will not be used. The number of columns added is the number of treated subjects that will be allowed to be discarded.

2.3. Cardinality matching

The third matching method we considered was cardinality matching [9]. Cardinality matching is an optimal matching method that maximizes the cardinality or size of the matched sample subject to constraints on covariate balance. As described in [9], the covariate balance constraints can be quite general. In their weakest form, they can require the means to be balanced (see [10] for details), but they can also require other forms of distributional balance such as fine balance [11] and strength-k balance [12]. To be precise, fine balance is a constraint on a nominal covariate that forces its marginal distributions to be identical, but does not require treated and control units to be matched within each of the categories of the nominal variable, as in exact matching (see Chapter 10 of [30] for details). Strength-k balancing is a stronger form of balance that forces low dimensional joint distributions of a nominal covariate to be identical; specifically, out of K nominal covariates, each of the $\binom{K}{l}$ possible interactions of covariates is finely balanced for all $l \leq k$, so the joint distributions of each of the $\sum_{l=1}^{k} \binom{K}{l}$ combinations of covariates is perfectly balanced. Clearly, strength-k balancing implies fine balance on each of the K nominal covariates. By imposing these constraints on covariate balance, cardinality matching directly balances the original covariates and does not require estimation of the propensity score or another summary of the covariates. Of course, stronger requirements on covariate balance tend to yield smaller matched samples.[‖]

From a computational standpoint, cardinality matching solves a linear integer programming problem, and, while a polynomial time algorithm to solve the cardinality matching problem has not been found, many instances of this problem can be solved in time comparable to the user time of the two previous methods. In addition, if finding an exact solution for an instance of the cardinality matching problem is too demanding, then it is possible to find an approximate solution by solving a relaxation of the integer programming problem [33]. This approximate solution may violate to some extent the covariate balancing constraints, but it will be found in polynomial time. In some settings, the balancing constraints can be formulated to require tighter balance than needed in view of the approximation.[**]

In our simulation study, we evaluated six types of covariate balance constraints with cardinality matching: (i) the widely used rule of absolute standardized differences in means, $|\hat{d}|$, smaller than 0.1, $|\hat{d}| < 0.1$; (ii) $|\hat{d}| < 0.01$; (iii) $|\hat{d}| < 0.001$; (iv) fine balance by discretizing continuous covariates into 10 categories; (v) both fine balance and $|\hat{d}| < 0.01$; (vi) and both fine balance and $|\hat{d}| < 0.001$. In some instances, we considered strength-2 balancing using 10 categories to discretize the continuous variables to balance their marginal distributions and 5 categories to balance the two-dimensional joint distributions together with $|\hat{d}| < 0.001$. We studied strength-2 balancing with 10 categories to balance both marginal and two-dimensional joints in the additional considerations section.

3. Simulation study design

3.1. Data generating mechanisms

Our simulation study design combines elements of the designs in [14] and [16]. Each simulated dataset consisted of $N = 250(1 + r)$ observations, where r is the number of controls available for each treated unit, so there were 250 treated units and $250r$ controls. The variables in each dataset consisted of three continuous outcomes and eight covariates, four of them continuous and four dichotomous. More specifically, the dichotomous covariates were two rare and two common Bernoulli random variables, all of them conditionally independent from each other and from the continuous covariates

[‖] In cardinality matching, finding the largest balanced matched sample is followed by re-pairing the treated and control units that constitute the matched sample to minimize their total sum of covariate distances. It is known that the heterogeneity of the matched pair differences in outcomes affects the sensitivity of results to biases due to unobserved covariates [6]. In cardinality matching, if the covariates used in the re-pairing are predictive of the outcome, this will reduce heterogeneity within matched groups and therefore sensitivity to biases due to unobserved covariates (see [31] for a general approach using the prognostic score, [32]).

[**] At the present, exact and approximate solutions to the cardinality matching problem can be found with the package `designmatch` for `R` [10, 33]. This package includes functions for the construction of matched samples that are balanced by design which can be used, among others, for matching in observational studies with treated and control units, as in this study, but also in settings with cases and controls (where the propensity score typically cannot be estimated) and under weaker identification assumptions with instrumental variables (e.g., [34]) and discontinuity designs [35].

given the treatment assignment indicator Z. The continuous covariates followed a multivariate normal distribution with covariance matrix $\boldsymbol{\Sigma}_t$ and mean vector $\boldsymbol{\mu}_t$ for the treated group, and $\boldsymbol{\Sigma}_c$ and $\boldsymbol{\mu}_c$ for the control group. The means of the eight covariates were selected in such a way that the true standardized differences in means, given by $d = \frac{\mu_t - \mu_c}{\sqrt{\frac{\sigma_t^2 + \sigma_c^2}{2}}}$ for the Normal variables and $d = \frac{p_t - p_c}{\sqrt{\frac{p_t(1-p_t)+p_c(1-p_c)}{2}}}$ for the Bernoulli random variables [20], were either 0.2 or 0.5.

For the covariance matrices of the Normal random variables, we considered three different scenarios of increasing complexity:

- Scenario 1, same variances in the treated and control groups with independent covariates in both groups;
- Scenario 2, different variances between treated and control groups with independent covariates in both groups;
- Scenario 3, different variances between treated and control groups with independent covariates in the control group and correlated covariates in the treatment group.

We generated the outcome Y from $Y = f(\boldsymbol{X}) + Z + \varepsilon$ where $\varepsilon \sim \mathcal{N}(0,4)$, and so here the true treatment effect is one. We considered three different forms of f:

- Linear, $f(x) = 3.5x_1 + 4.5x_3 + 1.5x_5 + 2.5x_7$;
- Additive, $f(x) = 3.5x_1 + 4.5x_3 + 1.5x_5 + 2.5x_7 + 2.5sign(x_1)|x_1|^{1/2} + 5.5x_3^2$;
- Additive with interactions, $f(x) = 3.5x_1 + 4.5x_3 + 1.5x_5 + 2.5x_7 + 2.5sign(x_1)|x_1|^{1/2} + 5.5x_3^2 + 2.5x_3x_7 - 4.5\left|x_1x_3^3\right|$.

The reader may notice that in each of these models only half of the covariates were included. This mimics the fact that in practice the investigator does not know exactly which covariates affect the outcome. As a consequence, in order to avoid discarding a potentially relevant covariate, it is often preferable to match for more rather than fewer covariates (see section 6.2 of [3]).

To assess the performance of the matching methods when there is different competition among the treated units for controls, we considered two values for the number of controls available per treated unit, $r = 1, 2$.[††] We generated a total of 1000 replications of each dataset and matched them with each method.

3.2. *Performance measures*

We compared the performance of the matching methods on the grounds of four criteria. First, we examined their ability to balance covariates. For this we used absolute standardized differences in means, variance ratios, Kolmogorov-Smirnov distances, and cross-match test statistics. More specifically, it has been said that absolute standardized differences in means smaller than 0.1 is evidence of covariate balance [36], but the fact that two covariates have distributions with similar first moments does not imply that differences in other moments do not introduce bias in the effect estimates (see sections 4 and 5; a related formal argument is proposition 4.2 of [37]). It is for this reason that we also calculated the variance ratio and the Kolmogorov-Smirnov distance between the empirical distributions of the continuous covariates across the treatment and control groups. Furthermore, to evaluate balance of the joint distributions we also calculated the standardized cross-match statistic [38]. Other interesting multivariate balance measures are discussed in [39, 40, 41, 42], but the cross-match statistic yields an exact, distribution free test for comparing two high-dimensional distributions, with close connections to the procedures used to construct a matched sample and a nice interpretation in terms of the propensity score [43]. To calculate this statistic we used the `R` package `crossmatch` [43] with the same propensity score distance used to match with nearest neighbor and optimal subset matching.

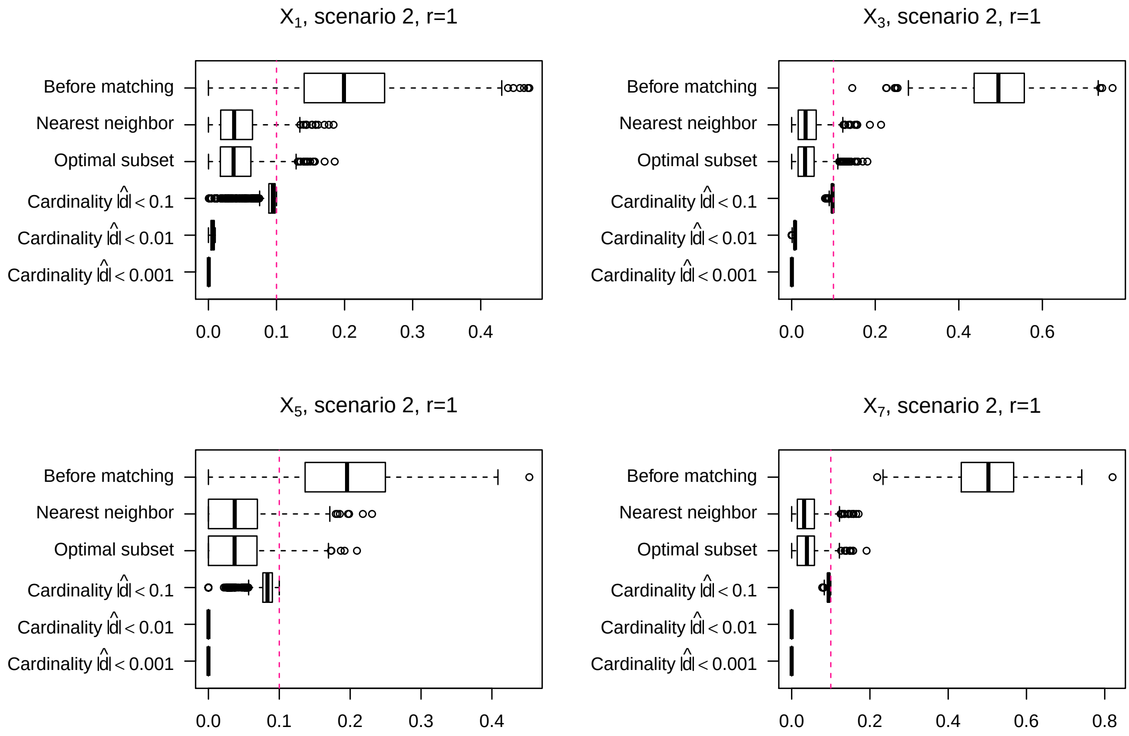
[††] We decided to fix the number of treated and control subjects, as in [14], instead of randomly assigning to treatment each unit by means of another Bernoulli random variable, as in [16], to have a fixed number of maximum possible pairs and be able to better compare the number of units matched across simulations.

The second criterion by which we compared the matching methods was sample size. For a given level of covariate balance (that is, for a given level of bias reduction), the largest matched sample should be preferred since this directly translates into more efficient estimates [44].

Next we used covariate distances. Though covariate distances play a secondary role in matching (they are often instrumental to achieve covariate balance), these distances can have an intrinsic importance depending on the statistical methods to be used after matching. If these methods explicitly use the matched-pair structure, then reducing covariate distances between matched pairs will increase efficiency and also reduce sensitivity to biases due to unobserved covariates [6]. To make cardinality matching comparable to the other matching methods in terms of distances, we re-paired the treated and control units initially selected by cardinality matching using optimal matching with a propensity score distance (as discussed in [9], re-pairing is actually the second stage of cardinality matching).

Finally, we compared the matching methods in terms of simple treatment effect estimates. We estimated the treatment effects using the difference in outcome means between the treatment and control groups and recorded the estimated bias and RMSE of these estimates. When looking at these values, it is important to keep in mind that, even if a matching method managed to perfectly balance the covariate distributions and consequently completely eliminate the bias due to observed covariates, the RMSE of the estimator would still be $\sqrt{\frac{2Var(\varepsilon)}{m}} = 2\sqrt{\frac{2}{m}}$, where m is the number of pairs selected when matching. This also explicitly shows why, for a fixed level of covariate balance, a larger sample size is preferable.

Figure 1. Boxplots of absolute standardized differences in means in scenario 2 when $r = 1$.



4. Results

4.1. Covariate balance

In this section we compare the matching methods in terms of covariate balance. Figure 1 summarizes the distribution of absolute standardized differences in means across simulated datasets in scenario 2 with $r = 1$ for four representative covariates. Other scenarios, group sizes ratios, and covariates presented a similar pattern. In the boxplots, we can see that the differences in means after matching with nearest neighbor and optimal subset matching varied to a fair extent from data set to data set, in many occasions being even over the 0.1 standard deviations threshold. On the other hand, with cardinality matching the absolute standardized differences in means were always smaller than 0.1, 0.01 or 0.001, as designed by the investigator. The evident performance difference takes place because with cardinality matching it is possible to finely-tune covariate balance adjustments, while with the other matching methods, covariate balance is attained indirectly, by matching on the propensity score and hoping that this will result in the desired covariate balance.

Table 1 shows summaries of balance for marginal distributions. In particular, it shows the averages of the variance ratios and Kolmogorov-Smirnov distances for each type of continuous covariate in every scenario when $r = 1$. The rest of the continuous covariates presented a similar pattern, as well as the case of $r = 2$. We observe that cardinality matching with fine balance constraints presented the variance ratios closest to 1 and the smallest K-S distance values in all the scenarios. This suggests that these are the best methods to balance the complete distribution of the covariates. This advantage was even more evident when the distributions before matching differed in more than just the mean (scenarios 2 and 3). When the variance ratio was originally one (scenario 1), all the resulting matches ended up with a variance ratio close to one. However, when the variances were different before matching, the only methods that corrected this were the ones involving fine balance; the rest of them kept the same ratio as before matching.

Table 1. Balance measures: variance ratio and Kolmogorov-Smirnov distance, $r = 1$

	Matching method	Scenario 1		Scenario 2		Scenario 3	
		$\frac{\sigma_t}{\sigma_c}$	K-S	$\frac{\sigma_t}{\sigma_c}$	K-S	$\frac{\sigma_t}{\sigma_c}$	K-S
X_1	Before matching	1.0021	0.1196	0.5066	0.1636	0.5050	0.1636
	Nearest neighbor	1.0065	0.0834	0.5128	0.1297	0.4999	0.1316
	Optimal subset	1.0039	0.0823	0.5125	0.1291	0.5002	0.1315
	Cardinality $\|\hat{d}\| < 0.1$	1.0054	0.0886	0.5095	0.1354	0.4967	0.1350
	Cardinality $\|\hat{d}\| < 0.01$	1.0056	0.0721	0.5110	0.1214	0.4937	0.1249
	Cardinality $\|\hat{d}\| < 0.001$	1.0052	0.0724	0.5103	0.1216	0.4930	0.1251
	Cardinality fine balance	1.0015	0.0574	0.9549	0.0654	0.9533	0.0655
	Cardinality $\|\hat{d}\| < 0.01 +$ fine balance	1.0008	0.0546	0.9543	0.0614	0.9528	0.0620
	Cardinality $\|\hat{d}\| < 0.001 +$ fine balance	1.0013	0.0542	0.9542	0.0613	0.9521	0.0615
X_3	Before matching	1.0065	0.2292	0.5020	0.2602	0.5035	0.2622
	Nearest neighbor	1.0175	0.0845	0.5450	0.1237	0.5136	0.1286
	Optimal subset	1.0142	0.0833	0.5434	0.1234	0.5138	0.1281
	Cardinality $\|\hat{d}\| < 0.1$	1.0063	0.0931	0.5200	0.1361	0.5033	0.1416
	Cardinality $\|\hat{d}\| < 0.01$	1.0110	0.0727	0.5263	0.1179	0.5036	0.1233
	Cardinality $\|\hat{d}\| < 0.001$	1.0113	0.0724	0.5261	0.1180	0.5029	0.1233
	Cardinality fine balance	0.9999	0.0595	0.9472	0.0684	0.9522	0.0670
	Cardinality $\|\hat{d}\| < 0.01 +$ fine balance	0.9999	0.0553	0.9471	0.0638	0.9512	0.0619
	Cardinality $\|\hat{d}\| < 0.001 +$ fine balance	0.9982	0.0550	0.9481	0.0635	0.9506	0.0618

Note: The values presented are the averages of the variance ratios and Kolmogorov-Smirnov distances observed with each method in each of the 1000 repetitions. Case $r = 2$ presents a similar pattern.

In short, the best marginal balance was obtained with cardinality matching when we combined fine balance and tight mean balance constraints. This resulted in a matching method that gave the best results on all marginal balance measures, in practically every scenario and control sample size. It is important to note that only requiring $|\hat{d}| < 0.1$ was the worst method in all cases and measures. In some sense, this calls into question the rule of absolute standardized differences in means smaller than 0.1. Optimal subset was better than nearest neighbor in attaining marginal balance, but both of them were worse than the other methods that required fine balance.

Table 2 compares the values observed for the cross-match statistic in all the simulation settings studied. For this measure we can distinguish three performance groups in all settings: distance driven methods, mean balance only methods, and fine balance methods. Fine balance methods produced the best results in terms of multivariate covariate balance, despite the fact that they did not explicitly include multidimensional balance constraints (such as strength-k balancing). The next best group was the mean balance only group, with values that were at least closer to zero than before matching. The worst group was the distance driven methods group. This group had the largest deviations from zero, in some occasions even larger than before matching, even though these methods focused on the same distance used to compute the cross-match statistic. Within this group, nearest neighbor was the best.

The ideal way to obtain k-dimensional balance would be to directly add this kind of constraints in the optimization problem, that is, to perform strength-k balancing. However, with 8 covariates, trying to obtain even a strength-2 match for all the observed covariates using fine categories for the continuous covariates can be very demanding and leave us with very few observations in this simulation setting. In a posterior section we discuss some results obtained when we perform strength-2 balancing on relevant covariates, in a situation in which we have more information on the relationship between the covariates and the outcome.

Table 2. Cross-match test statistic

	Matching method	Scenario 1	Scenario 2	Scenario 3
$r = 1$	Before matching	-3.8466	-3.8693	-3.5923
	Nearest neighbor	2.4963	2.1126	2.2819
	Optimal subset	4.1896	4.2104	4.1977
	Cardinality $\|\hat{d}\| < 0.1$	-1.2758	-1.2974	-1.1795
	Cardinality $\|\hat{d}\| < 0.01$	-1.5434	-1.4843	-1.4886
	Cardinality $\|\hat{d}\| < 0.001$	-1.5388	-1.4903	-1.4661
	Cardinality fine balance	-0.1898	-0.0293	0.0150
	Cardinality $\|\hat{d}\| < 0.01 +$ fine balance	-0.1922	0.0131	-0.0058
	Cardinality $\|\hat{d}\| < 0.001 +$ fine balance	-0.1665	-0.0352	-0.0086
$r = 2$	Before matching	-4.3518	-4.1156	-3.9883
	Nearest neighbor	4.4745	3.9573	4.3405
	Optimal subset	5.4422	5.0051	5.2760
	Cardinality $\|\hat{d}\| < 0.1$	-2.0475	-1.1871	-1.6029
	Cardinality $\|\hat{d}\| < 0.01$	-2.3067	-1.7510	-2.0163
	Cardinality $\|\hat{d}\| < 0.001$	-2.2674	-1.7854	-2.0042
	Cardinality fine balance	-0.6880	-0.1225	-0.2265
	Cardinality $\|\hat{d}\| < 0.01 +$ fine balance	-0.6204	-0.1633	-0.2100
	Cardinality $\|\hat{d}\| < 0.001 +$ fine balance	-0.6092	-0.1872	-0.2291

Note: The values presented are the averages of the standardized cross-match statistics observed with each method in each of the 1,000 repetitions. The values are standardized by subtracting the expectation under the null hypothesis of equal distribution in SE units.

4.2. Sample sizes

Table 3 presents the average number of pairs matched with each method on every scenario and control sample size. Naturally, when there were more controls available per treated subject, every method was able to obtain a larger sample size. In our example, the number of matched pairs when $r = 2$ was between approximately 30% to 45% larger than the number of pairs matched when $r = 1$. The comparison among matching methods was similar for both values of r.

Cardinality matching with $|\hat{d}| < 0.1$ was always the matching method that produced the largest sample sizes, followed by the other mean balance methods, decreasing the sample size as the mean balance was tightened. For the other matching methods, the comparison depended on the type of imbalances in the covariate distributions before matching. When they differed only by their means (scenario 1), cardinality matching with fine balance methods kept more or a similar amount of observations than the distance driven methods. This changed when the observed covariates were more imbalanced before matching (scenario 2 and 3). In these cases, the latter methods had larger sample sizes (between 8% and 20% higher) than the fine balance methods. This happened because the fine balance methods had to discard more observations in order to correct important distributional imbalances which, as studied on the covariance balance section, the distance driven methods failed to accomplish. Again, in all scenarios, the sample size for cardinality matching with fine balance was smaller as tighter mean balance was required. Also, in every case, optimal subset matched more pairs than nearest neighbor, although the difference was small (less than 2%).

As we have seen, direct comparison of the sizes of the subsets selected by each method is rather unfair. A more valuable comparison would be to observe the sample sizes at a certain covariate balance level. This balance level cannot be set beforehand with nearest neighbor and optimal subset matching. Nevertheless, we can notice that cardinality matching with $|\hat{d}| < 0.01$ and $|\hat{d}| < 0.001$ produced a similar but better covariance balance than those methods and they provided sample sizes at least 10% larger than those observed with nearest neighbor or optimal subset matching. This suggests that at certain level of covariate balance, the sizes of the subsets that result from cardinality matching are larger than with the other methods.

Table 3. Sample size

	Matching method	Scenario 1	Scenario 2	Scenario 3
$r = 1$	Nearest neighbor	145.39	144.92	148.84
	Optimal subset	148.16	147.71	151.44
	Cardinality $|\hat{d}| < 0.1$	182.88	182.80	184.89
	Cardinality $|\hat{d}| < 0.01$	164.50	164.46	167.82
	Cardinality $|\hat{d}| < 0.001$	163.44	163.44	167.01
	Cardinality fine balance	149.72	126.40	128.25
	Cardinality $|\hat{d}| < 0.01$ + fine balance	148.79	125.40	127.31
	Cardinality $|\hat{d}| < 0.001$ + fine balance*	147.30	124.36	126.29
$r = 2$	Nearest neighbor	192.49	200.13	199.95
	Optimal subset	195.80	203.42	203.12
	Cardinality $|\hat{d}| < 0.1$	241.83	243.95	244.12
	Cardinality $|\hat{d}| < 0.01$	222.51	226.83	228.41
	Cardinality $|\hat{d}| < 0.001$*	220.57	224.90	226.73
	Cardinality fine balance	208.11	185.24	186.89
	Cardinality $|\hat{d}| < 0.01$ + fine balance	207.12	184.13	185.83
	Cardinality $|\hat{d}| < 0.001$ + fine balance *	205.71	182.93	184.69

Note: The values presented are the averages of the sample sizes observed with each method in each of the 1,000 repetitions.

* For some of the simulated datasets, these methods did not find a solution in the allotted time. The averages do not include those cases.

4.3. *Distances*

Table 4 shows the average propensity score distance between matched pairs obtained by each method. As expected, in terms of distances, the method that performed the best was optimal subset matching, followed by nearest neighbor matching. Optimal subset matching is designed precisely to minimize the global distance between matched pairs while matching as many pairs as possible. In contrast, cardinality matching maximizes the number of observations that satisfy some balance constraints. This explains the large difference between nearest neighbor matching and cardinality matching with fine balance and $|\hat{d}| < 0.001$ constraints, which was the next best method. The average distance for the latter was between 3 to 8 times as large as nearest neighbor, depending on the scenario and the control sample size. The average distance increased as the distributional constraints to which the cardinality matching method was subject to were relaxed, resulting in the following order from best to worst: fine balance and $|\hat{d}| < 0.01$, fine balance, $|\hat{d}| < 0.001$, $|\hat{d}| < 0.01$, and, lastly, $|\hat{d}| < 0.1$. The results hold for all the scenarios and for both control sample sizes analyzed.

Table 4. Average propensity score distance

	Matching method	Scenario 1	Scenario 2	Scenario 3
$r = 1$	Nearest neighbor	0.0377	0.0405	0.0367
	Optimal subset	0.0241	0.0242	0.0223
	Cardinality $\|\hat{d}\| < 0.1$	0.3547	0.3606	0.3258
	Cardinality $\|\hat{d}\| < 0.01$	0.2444	0.2494	0.2315
	Cardinality $\|\hat{d}\| < 0.001$	0.2401	0.2450	0.2281
	Cardinality fine balance	0.1417	0.1250	0.1154
	Cardinality $\|\hat{d}\| < 0.01 +$ fine balance	0.1361	0.1208	0.1129
	Cardinality $\|\hat{d}\| < 0.001 +$ fine balance	0.1305	0.1171	0.1114
$r = 2$	Nearest neighbor	0.0238	0.0251	0.0227
	Optimal subset	0.0192	0.0196	0.0180
	Cardinality $\|\hat{d}\| < 0.1$	0.4341	0.3442	0.3612
	Cardinality $\|\hat{d}\| < 0.01$	0.3196	0.2570	0.2826
	Cardinality $\|\hat{d}\| < 0.001$	0.3117	0.2509	0.2764
	Cardinality fine balance	0.2064	0.1133	0.1268
	Cardinality $\|\hat{d}\| < 0.01 +$ fine balance	0.2023	0.1099	0.8898
	Cardinality $\|\hat{d}\| < 0.001 +$ fine balance	0.1963	0.1082	0.1241

Note: The values presented are the averages of the average propensity score distance observed with each method in each of the 1000 repetitions.

4.4. *Treatment effect estimates*

We now study which method provided the best treatment effect estimates for each form of relationship between the outcome and the observed covariates. Table 5 shows the bias and root mean square error (RMSE) for each method when $r = 1$. Results for $r = 2$ were slightly better for all the matching methods than when $r = 1$; however, the comparison among methods was similar.

When the outcome is a linear combination of the observed covariates, balancing the means of the covariates suffices to remove systematic biases in the treatment effect estimates. This means that the smaller the absolute standardized differences in means are, the less biased is the treatment effect estimate. For instance, if we compare Table 5 with Figure 1, we notice that the methods with lower and less variable absolute standardized differences in means were the ones that showed a closer and more precise estimate. In this case, cardinality matching with $|\hat{d}| < 0.001$ had the best results, followed by $|\hat{d}| < 0.01$. In some instances, these two showed a greater bias than those obtained by distance driven methods;

nonetheless, their variability was lower and consequently the RMSE was, in some cases, half the RMSE obtained with the distance driven methods. In general, when the outcome is linear, it appears to be advisable to balance the means as closely as possible, as long as enough observations are matched. Note that in our setting we were able to tighten the balance constraints from $|\hat{d}| < 0.01$ to $|\hat{d}| < 0.001$ with practically no reduction in sample size.

Table 5. Treatment effect estimation performance, $r = 1$

	Matching method	Scenario 1		Scenario 2		Scenario 3	
		Bias	RMSE	Bias	RMSE	Bias	RMSE
Linear	Before matching	3.6210	3.6649	4.2477	4.2979	4.2741	4.3333
	Nearest neighbor	0.0416	0.3601	-0.0058	0.4177	0.0506	0.4090
	Optimal subset	0.0322	0.3414	0.0072	0.4053	0.0563	0.3887
	Cardinality $\|\hat{d}\| < 0.1$	0.8790	0.9086	1.0269	1.0537	0.9627	1.0062
	Cardinality $\|\hat{d}\| < 0.01$	0.0531	0.2299	0.0535	0.2244	0.0599	0.2302
	Cardinality $\|\hat{d}\| < 0.001$	0.0063	0.2233	-0.0042	0.2164	0.0115	0.2178
	Cardinality fine balance	0.0857	0.2762	0.1074	0.3203	0.1018	0.3241
	Card. $\|\hat{d}\| < 0.01$ + fine balance	0.0248	0.2340	0.0219	0.2521	0.0285	0.2649
	Card. $\|\hat{d}\| < 0.001$ + fine balance	0.0038	0.2163	0.0010	0.2220	0.0094	0.2343
	Card. $\|\hat{d}\| < 0.01$ + strength-2*	0.2002	0.3887	0.1958	0.4347	0.1803	0.4469
Additive	Before matching	5.4213	5.5432	1.3089	1.9695	1.3327	2.0644
	Nearest neighbor	0.0777	1.0287	-4.4930	4.7318	-4.8763	5.0937
	Optimal subset	0.0479	0.9863	-4.4849	4.7225	-4.8469	5.0582
	Cardinality $\|\hat{d}\| < 0.1$	1.3111	1.5378	-3.0667	3.3076	-3.4446	3.6783
	Cardinality $\|\hat{d}\| < 0.01$	0.0790	0.8214	-4.5232	4.6963	-4.8355	5.0088
	Cardinality $\|\hat{d}\| < 0.001$	0.0051	0.8082	-4.6309	4.8011	-4.9353	5.1064
	Cardinality fine balance	0.1312	0.4070	-0.1644	0.5608	-0.1507	0.5695
	Card. $\|\hat{d}\| < 0.01$ + fine balance	0.0323	0.3334	-0.3241	0.5544	-0.2963	0.5501
	Card. $\|\hat{d}\| < 0.001$ + fine balance	-0.0123	0.2981	-0.3635	0.5173	-0.3419	0.5217
	Card. $\|\hat{d}\| < 0.01$ + strength-2*	0.2807	0.5872	-0.0950	0.7035	-0.1518	0.7392
With interactions	Before matching	3.9501	4.1854	15.8023	16.2399	14.7405	15.2330
	Nearest neighbor	0.0966	1.4362	10.7838	11.7255	10.7544	11.6885
	Optimal subset	0.0829	1.4655	10.7368	11.6525	10.7005	11.6010
	Cardinality $\|\hat{d}\| < 0.1$	1.1222	1.6874	11.5650	12.2497	11.2694	11.9708
	Cardinality $\|\hat{d}\| < 0.01$	0.1338	1.2818	10.5861	11.3906	10.5498	11.3333
	Cardinality $\|\hat{d}\| < 0.001$	0.0822	1.2882	10.5860	11.4048	10.5906	11.4058
	Cardinality fine balance	0.1022	1.0648	0.6323	2.3448	-0.8431	2.4767
	Card. $\|\hat{d}\| < 0.01$ + fine balance	0.0246	1.0653	0.6041	2.3249	-0.8564	2.5347
	Card. $\|\hat{d}\| < 0.001$ + fine balance	0.0219	0.9926	0.5875	2.1095	-0.8707	2.3721
	Card. $\|\hat{d}\| < 0.01$ + strength-2*	0.2649	0.7434	0.7491	1.6680	0.5846	1.4790

Note: The bias and RMSE observed with each method with $r = 2$ present a similar pattern as shown in this table.

* Matches were obtained using the approximation algorithm in `designmatch` [33].

When the outcome is not a linear combination of the observed covariates, which will most likely be the case in practice, it appears that a stronger form of covariate balance is needed. In this case, balancing only the means is not enough to remove systematic biases, unless the distributions of the covariates only differ by their means before matching (like in scenario 1) and the rest of the distributions remain balanced after matching. In our simulation we can see that even though the bias of cardinality matching with only $|\hat{d}| < 0.001$ was among the smallest, its RMSE was more than 2.5 times the RMSE of cardinality matching that in addition to $|\hat{d}| < 0.001$ required fine balance. This occurred because the magnitude of the differences of the complete matched distributions varied from dataset to dataset, which translated into differences in the nonlinear part of the outcome and therefore in the treatment effect estimate.

When the covariate distributions differed in more ways than in their means (scenarios 2 and 3), the only methods that corrected these differences were the ones involving fine balance and strength-2 matching. It is clear from Table 5 that these distributional differences can affect to a great extent the bias and RMSE of the treatment effect estimates. This depends on the importance and level of nonlinearity of the nonlinear part of the outcome, but in our simulation we observed RMSEs 5 to 10 times higher on the rest of the methods compared to cardinality matching with fine balance and strength-2.

In general, when the outcome is an additive function without interactions, the ideal way of matching is with fine balance plus tight mean balance constraints, whereas when the outcome is an additive function with interactions, it is best to match with strength-k balancing, again with tight mean balance. As we mentioned earlier, stronger requirements on covariate balance tend to yield smaller matched samples, and thus, matching with strength-k can be very demanding in terms of sample size. Here, the strength-2 balancing considered for all the possible pairs of the eight covariates resulted in matched samples that ranged in size between 35% and 40% of the original number of treated units available when $r = 1$ and between 57% and 62% when $r = 2$. However, if we had used 10 categories to balance both marginal and two-dimensional joint distributions, sample sizes could have been as small as 5 observations, which would be unacceptable. In practice, the decision to require stronger forms of covariate balance must be weighed against the resulting sample size.

In summary, if the outcome is known to be a linear function of the observed covariates, balancing the means would be enough. A good way to do this is to use cardinality matching to directly balance the means as closely as the data allows without sacrificing too much sample size. However, it is quite unlikely that this will be the case, so it is important to always consider the possibility that there are nonlinearities. Not doing this can result in extremely biased estimations. Thus, if the treatment effect is estimated with no further covariate adjustments, it is recommended to use cardinality matching with both fine balance and tight mean balance constraints if the outcome model is believed to be additive, and with strength-k balancing and tight mean balance constraints if the model is conjectured to be additive with interactions.

5. Additional considerations

5.1. Incorporating prior knowledge about the relationship between the observed covariates and the outcome

In this section we analyze the effect estimates when the investigator has different levels of knowledge about the true functional form that relates the outcome and the covariates, and this knowledge is included in the matching process by imposing different forms of covariate balance. Here we focus on cardinality matching since this method gives the investigator the flexibility to directly obtain different forms of covariate balance. From the previous simulation results section, for each functional form we selected the form of covariate balance that provided the best treatment effect estimates and compared these estimates with the ones obtained when different levels of knowledge were incorporated into the matching process. These levels of knowledge are referred to as “no,” “weak,” “strong,” and “complete knowledge” about the true function that relates the covariates to the outcome.

Specifically, no knowledge refers to the situation where the investigator does not know which covariates affect the outcome nor in which way (this is, linearly, nonlinearly without interactions, or nonlinearly allowing for interactions between the covariates). Weak knowledge refers to the case where the investigator knows which covariates are relevant in terms of the outcome, but does not know if they relate to the outcome linearly, nonlinearly without interactions, or nonlinearly with interactions. Strong knowledge is the case where one knows not only which covariates are relevant, but also, to some extent, the role they play in the outcome model. For example, this would be the case if one knew which covariates were relevant in a non-linear way and which covariates interacted between each other, but without specifically knowing the non-linear functional form or how they interacted. Finally, complete knowledge refers to the utopic situation where the investigator knows the true functional form of the model that relates the covariates and the outcome. This case is only considered as a benchmark to evaluate the rest of the matches. The forms of covariate balance considered for each of these situations are described in Table 6.

Table 7 and Table 8 summarize the main results in terms of treatment effect estimation and sample sizes. The results show that, when the outcome is linear, having more previous knowledge does not necessarily translate into better treatment effect estimation. However, having more information allows the distributional balance constraints to be relaxed as needed, obtaining, as a consequence, larger sample sizes and therefore more precise estimates. When the outcome is additive, again, the RMSEs of the treatment effect estimate are close between each other when there is no complete knowledge. Nonetheless, in our case, these RMSEs were between 50% larger to more than twice the size obtained when there was complete knowledge. This is due to the fact that continuous covariates were approximately balanced by balancing their categorized versions instead of being perfectly balanced. When there are interactions, it is possible to appreciate the great advantage of reducing the number of covariates we match on, especially when only strength-2 balancing of a coarsely discretized version of the continuous covariates is feasible, which was the case in this simulation setting. Being able to perform strength-2 balancing with finer categories on the variables that interact, significantly improved the estimation of the treatment effect compared to the performance when only fine balance or a coarse version of strength-2 was obtained.

In short, if we want to minimize our estimates' dependence on model assumptions, the most appropriate choice is to match with the strongest constraints, that is, to require tight mean balance, fine balance, and strength-2 balance. Doing this will result in smaller sample sizes, but the consequences of this will depend on the original number of observations available. However, if relevant covariate distributions are not properly balanced, the matching process will fail to remove an important part of the original bias. If the available data does not allow the researcher to obtain this degree of balance, some relaxation of constraints will be needed. It would be ideal if the researcher had more information about the relationship between the outcome and the covariates in order to use it to determine which constraints to relax. For instance, this knowledge could be obtained by splitting the sample into a small planification sample to learn features of the outcome model (for instance, by using LASSO regression with the original covariates and transformations of them) and a larger analysis sample to conduct the actual matching and outcome analyses (see [45, 46] for related ideas).

5.2. *Larger number of covariates*

We explored the performance of cardinality matching with a larger number of covariates using approximation algorithms. With 50 covariates, we were not able to find an exact solution within a one-hour time window, whereas by using the approximation algorithm in `designmatch` [33] we typically found a solution in a few minutes. This approximate solution occasionally violated some of the covariate balancing constraints, but the resulting balance was systematically better than the one with the distance-based methods. In general, the outcome results were qualitatively similar than with 8 covariates. We also explored the case with 100 covariates, but we were not able to find a solution because of memory constraints. In general, for matching problems with large number of covariates, a practical way to proceed is to use the approximate algorithm in `designmatch` and then either tighten the balancing requirements if the balancing constraints are violated by too much, or use this approximate solution as a "warm start" to find an exact solution in shorter amount of time. Broadly, the use of approximation algorithms has not been explored much in matching in observational studies and it is an interesting area of research.

5.3. *Heterogenous effects*

When the effects are homogeneous, as in our simulation study, the average treatment effect on the treated units is the same as the average treatment effect on the *matched* treated units, but this is not necessarily the case when the effects are heterogeneous. When the effects are heterogeneous, if the estimand is the average treatment effect on the matched treated

Table 6. Forms of covariate balance considered for different levels of knowledge about the true outcome model

	Outcome model		
	Linear	Additive	Additive with interactions
Knowledge			
No	$\lvert\hat{d}\rvert < 0.001$ for $X_1, X_2, ..., X_8$;	$\lvert\hat{d}\rvert < 0.001 +$ f.b. for $X_1, X_2, ..., X_8$	$\lvert\hat{d}\rvert < 0.001 +$ f.b. for $X_1, X_2, ..., X_8$
	$\lvert\hat{d}\rvert < 0.001 +$ f.b. for $X_1, X_2, ..., X_8$		$\lvert\hat{d}\rvert < 0.001 +$ s.-2 for $X_1, X_2, ..., X_8$*
Weak	$\lvert\hat{d}\rvert < 0.001$ for X_1, X_3, X_5, X_7;	$\lvert\hat{d}\rvert < 0.001 +$ f.b. for X_1, X_3, X_5, X_7;	$\lvert\hat{d}\rvert < 0.001 +$ f.b. for X_1, X_3, X_5, X_7;
	$\lvert\hat{d}\rvert < 0.001 +$ f.b. for X_1, X_3, X_5, X_7;	$\lvert\hat{d}\rvert < 0.001 +$ s.-2 for X_1, X_3, X_5, X_7	$\lvert\hat{d}\rvert < 0.001 +$ s.-2 for X_1, X_3, X_5, X_7
	$\lvert\hat{d}\rvert < 0.001 +$ s.-2 for X_1, X_3, X_5, X_7		
Strong	$\lvert\hat{d}\rvert < 0.001$ for X_1, X_3, X_5, X_7	$\lvert\hat{d}\rvert < 0.001$ for $X_5, X_7 + \lvert\hat{d}\rvert < 0.001$ and f.b. for X_1, X_3	$\lvert\hat{d}\rvert < 0.001$ for $X_1, X_3, X_5, X_7 +$ f.b. for $X_1, X_3 +$ s.-2 for X_1, X_3 and X_3, X_7
Complete	$\lvert\hat{d}\rvert < 0.001$ for X_1, X_3, X_5, X_7	$\lvert\hat{d}\rvert < 0.001$ for X_1, X_3, X_5, X_7, $sign(X_1)\lvert X_1\rvert^{1/2}, X_3^2$	$\lvert\hat{d}\rvert < 0.001$ for X_1, X_3, X_5, X_7, $sign(X_1)\lvert X_1\rvert^{1/2}, X_3^2, X_3X_7, \lvert X_1X_3^3\rvert$

Note: $|\hat{d}| < 0.001$ denotes absolute differences in means smaller than 0.001 standard deviations; f.b. stands for fine balance, and s.-2 for strength-2 balancing. In this way, for example, when there is no knowledge about the form of the true outcome model, and this model is linear (top left corner of the table), we imposed two forms of covariate balance: absolute differences in means smaller than 0.001 standard deviations for all the covariates (this is, $|\hat{d}| < 0.001$ for $X_1, X_2, ..., X_8$), and both absolute differences in means smaller than 0.001 standard deviations and fine balance for all the covariates ($|\hat{d}| < 0.001 +$ f.b. for $X_1, X_2, ..., X_8$).

* Using the approximation algorithm in `designmatch` [33] with 5 categories to balance two-dimensional joints.

Table 7. Treatment effect estimation performance at different levels of previous knowledge, $r = 1$

	Matching method	Scenario 1		Scenario 2		Scenario 3	
		Bias	RMSE	Bias	RMSE	Bias	RMSE
Linear	Before matching	3.6210	3.6649	4.2477	4.2979	4.2741	4.3333
	All: Cardinality $\|\hat{d}\| < 0.001$	0.0063	0.2233	-0.0042	0.2164	0.0115	0.2178
	All: Cardinality $\|\hat{d}\| < 0.001$ + fine balance	0.0038	0.2163	0.0010	0.2220	0.0094	0.2343
	W/S/C: Cardinality $\|\hat{d}\| < 0.001$	0.0061	0.2082	-0.0132	0.2047	0.0051	0.2053
	Weak: Cardinality $\|\hat{d}\| < 0.001$ + fine balance	0.0052	0.2107	-0.0088	0.2178	-0.0015	0.2232
	Weak: Cardinality $\|\hat{d}\| < 0.001$ + strength-2	0.0072	0.2360	-0.0065	0.2383	-0.0010	0.2450
Additive	Before matching	5.4213	5.5432	1.3089	1.9695	1.3327	2.0644
	All: Cardinality $\|\hat{d}\| < 0.001$ + fine balance	-0.0123	0.2981	-0.3635	0.5173	-0.3419	0.5217
	Weak: Cardinality $\|\hat{d}\| < 0.001$ + fine balance	-0.0021	0.2909	-0.3376	0.5060	-0.3423	0.5108
	Weak: Cardinality $\|\hat{d}\| < 0.001$ + strength-2	0.0061	0.3125	-0.3046	0.5066	-0.3029	0.4993
	Strong: Cardinality $\|\hat{d}\| < 0.001$ + fine balance	-0.0016	0.2967	-0.3461	0.5124	-0.3417	0.5130
	Complete: $\|\hat{d}\| < 0.001$	0.0046	0.2063	-0.0156	0.2099	0.0022	0.2122
Additive with interactions	Before matching	3.9501	4.1854	15.8023	16.2399	14.7405	15.233
	All: Cardinality $\|\hat{d}\| < 0.001$ + fine balance	0.0219	0.9926	0.5875	2.1095	-0.8707	2.3721
	All: Cardinality $\|\hat{d}\| < 0.01$ + strength-2*	0.2649	0.7434	0.7491	1.6680	0.5846	1.4790
	Weak: Cardinality $\|\hat{d}\| < 0.001$ + fine balance	0.0208	0.9303	0.5592	1.9611	-0.6769	2.1902
	Weak: Cardinality $\|\hat{d}\| < 0.001$ + strength-2	-0.0001	0.3863	0.3396	0.8675	0.2883	0.9219
	Strong: Cardinality $\|\hat{d}\| < 0.001$ + fine balance	0.0169	0.9345	0.5498	1.9155	-0.6471	2.1735
	Strong: Cardinality $\|\hat{d}\| < 0.001$ + strength-2	0.0006	0.3867	0.3586	0.8644	0.3445	0.9356
	Complete: $\|\hat{d}\| < 0.001$	0.0040	0.2099	-0.0049	0.2130	0.0051	0.2124

Note: The bias and RMSE observed with each method and level of knowledge with $r = 2$ present a similar pattern as the shown in this table.
* Using the approximation algorithm in `designmatch` [33] with 5 categories to balance two-dimensional joints.

Table 8. Sample sizes when matching with different levels of previous knowledge

	Matching method	Scenario 1		Scenario 2		Scenario 3	
		$r = 1$	$r = 2$	$r = 1$	$r = 2$	$r = 1$	$r = 2$
Linear	All: Cardinality $\lvert\hat{d}\rvert < 0.001$	163.44	220.57	163.44	224.90	167.01	226.73
	All: Cardinality $\lvert\hat{d}\rvert < 0.001$ + fine balance all	147.30	205.71	124.36	182.93	126.29	184.69
	W/S/C: Cardinality $\lvert\hat{d}\rvert < 0.001$	186.22	244.32	186.24	244.73	186.98	244.59
	Weak: Cardinality $\lvert\hat{d}\rvert < 0.001$ + fine balance	178.36	235.09	164.35	230.52	164.52	230.91
	Weak: Cardinality $\lvert\hat{d}\rvert < 0.001$ + strength-2	144.38	199.46	136.01	199.14	133.86	197.40
Additive	All: Cardinality $\lvert\hat{d}\rvert < 0.001$ + fine balance	147.30	205.71	124.36	182.93	126.29	184.69
	Weak: Cardinality $\lvert\hat{d}\rvert < 0.001$ + fine balance	178.36	235.09	164.35	230.52	164.52	230.91
	Weak: Cardinality $\lvert\hat{d}\rvert < 0.001$ + strength-2	144.38	199.46	136.01	199.14	133.86	197.40
	Strong: Cardinality $\lvert\hat{d}\rvert < 0.001$ + fine balance	178.36	235.09	164.35	230.52	164.52	230.91
	Complete: Cardinality $\lvert\hat{d}\rvert < 0.001$	185.75	243.41	178.56	242.80	179.16	242.68
Additive with interactions	All: Cardinality $\lvert\hat{d}\rvert < 0.001$ + fine balance	147.30	205.71	124.36	182.93	126.29	184.69
	All: Cardinality $\lvert\hat{d}\rvert < 0.01$ + strength-2*	100.20	154.00	89.69	147.60	86.68	142.80
	Weak: Cardinality $\lvert\hat{d}\rvert < 0.001$ + fine balance	178.36	235.09	164.35	230.52	164.52	230.91
	Weak: Cardinality $\lvert\hat{d}\rvert < 0.001$ + strength-2	144.38	199.46	136.01	199.14	133.86	197.40
	Strong: Cardinality $\lvert\hat{d}\rvert < 0.001$ + fine balance	178.35	235.09	164.35	230.51	164.52	230.91
	Strong: Cardinality $\lvert\hat{d}\rvert < 0.001$ + strength-2	151.84	208.50	142.63	207.06	140.51	205.10
	Complete: Cardinality $\lvert\hat{d}\rvert < 0.001$	184.46	237.99	176.05	237.37	177.31	238.03

Note: Some values are repeated because some matchings are the same for different functional forms.
* Using the approximation algorithm in `designmatch` [33] with 5 categories to balance two-dimensional joints.

units, then all our results carry over, as the only source of bias is that of imbalances in the functions of the covariates that intervene in the true outcome model. If the estimand is the average treatment effect on the treated units, and all the treated units cannot be matched due to limited overlap in covariate distributions between the matched groups, then, in addition to the previous bias due to imbalances, there will be another source of bias due to incomplete matching [47]. In settings where there is limited overlap in covariate distributions, we find more meaningful to target the average treatment effect on the matched treated units, as any estimate of the average treatment effect on the treated units will rely to some extent on extrapolation. In view of this limitation imposed by the data, one way to make progress without making further modeling assumptions is by describing both the matched and unmatched samples [48]. This gives a basic understanding of the population to which the results of the matched analysis can be generalized in principle (see [49] and [50] for related methods).

5.4. *Correct specification of the propensity score model*

Throughout we have estimated the propensity score using logistic regression, including all the covariates as linear terms in the propensity score model [16, 17, 18]. We also explored the changes that would be observed if the propensity score was estimated in a more flexible way, for example, by including higher order and interaction terms of the covariates that intervene in the true outcome and propensity score models. When the true outcome model was additive, it made a substantial difference to include these terms in the propensity score model, as both bias and RMSE were considerably reduced in scenarios 2 and 3 for the distance-based methods. Something similar happened when the true outcome model was additive with interactions. In some of these instances, when the propensity score model was specified in accordance with the true outcome and propensity score models, bias was the lowest with the distance-based methods, however the estimates were quite variable and cardinality matching with fine balance still achieved lower RMSEs. In practice, a good alternative to distance driven matching methods may be to estimate the propensity score using a more flexible approach than logistic regression, for example, by using ensemble methods as in [51].

5.5. *Limited overlap in covariate distributions*

Assessing limited overlap or lack of common support in covariate distributions is a common practice in observational studies. Its goal is to avoid extrapolating or fabricating results from models that assume specific functional forms (see Section 18.2 of [30] and Chapter 14 of [52] for related discussions). This assessment is typically done, first, by trimming the sample on the propensity score, and second, by checking balance (see, for instance, [53] and [54]). In contrast, cardinality matching directly trims the sample by selecting the largest matched sample that satisfies the investigator's requirements for covariate balance. Certainly, if there is no overlap in a given covariate, then this matched sample will be empty. In a sense, since cardinality matching maximizes the size of the matched sample that is balanced, the sample it finds constitutes the portion of the data set of "maximal" overlap of the covariates distributions given the requirements for covariate balance.

5.6. *Sensitivity to hidden biases*

In standard practice, the construction of a pair-matched sample is done in a single step, by selecting pairs of treated and control units that are similar in terms of a summary of the covariates and hoping that these paired groups will be balanced on aggregate, in terms of each of the covariates. By contrast, in cardinality matching the tasks of balancing and pairing are separated into two steps. First, the method finds the largest pair-matched sample that is balanced, and then, with the balanced sample in hand, it re-pairs the treated and control units minimizing the total sum of distances in some of the covariates. It is known that reducing heterogeneity in pair differences in outcomes results in reduced sensitivity to hidden bias [6]. Therefore, if the covariates used in re-pairing are strong predictors of the outcome, this second stage will results in reduced sensitivity. This is something that can be exploited in practice either by relying on substantive knowledge of which covariates are predictive of the outcome or by learning them from the data itself (see [31] for an interesting related method).

5.7. Exploring the trade-off between covariate balance and sample size

With any matching method that selects a subset of treated and control observations there is a tension between covariate balance and sample size; this is, a bias-variance trade-off between removing biases due to imbalances in observed covariates and using a larger matched sample to thereby reduce variance. In this study, we have called into question the widely used rule of thumb of balancing covariates so that their differences in means are not greater than 0.1 standard deviations, and gave broad recommendations to balance covariates under general outcome models. Of course, the applicability of these recommendations will depend on the available data and the resulting sample size after matching. To select a particular balance-size matched design, one way to proceed in the spirit of [55] is to plot the covariate balance-sample size pairs in a two-dimensional plot to explore this trade-off and select a design in the plot. Since this selection does not require the outcomes, it does not affect the objectivity of the study nor the validity of the statistical tests [56]. As we mentioned in the introduction, how to optimally balance covariates is an open question to which this simulation study has gave some answers, and which we believe should ultimately be addressed with formal statistical theory.

6. Summary and remarks

We presented a Monte Carlo simulation study of three multivariate matching methods that select a subset of treatment and control observations: the widely used nearest neighbor matching with propensity score calipers, and the more recently proposed, optimal subset matching and cardinality matching. We evaluated the performance of these methods according to four different criteria: covariate balance, sample size, covariate distances between matched pairs, and treatment effect estimates. The main findings are the following.

In terms of covariate balance, cardinality matching had the best performance among the three methods. As shown in Figure 1, cardinality matching gives the investigator a precise degree of control over covariate balance adjustments. For example, the investigator can require the absolute standardized differences in means to be smaller than 0.1, 0.01, 0.001, and so on, and at the same time directly balance marginal and k-way joint distributions via fine balance and strength-k balancing. In principle, if the investigator assigns more importance to some covariates than others, he or she can balance these covariates more tightly by imposing stronger mean balance or distributional balance constraints on them. Unlike most matching methods, with cardinality matching the propensity score is not needed to balance the covariates because it directly balances the original covariates; however, with cardinality matching the propensity score may as well be balanced as an additional covariate.

In terms of the size of the matched samples, the results show that, for a given level of covariate balance (e.g., for absolute standardized differences in means smaller than 0.01), cardinality matching systematically selects a larger subset of the observations. Certainly, for stronger forms of covariate balance, the size of the matched sample will be smaller.

In terms of covariate distances, optimal subset matching exhibits the best performance among the three methods. While often considered an instrumental objective in order to balance covariates, reducing covariate distances between matched pairs can be an objective per se because it results into reduced heterogeneity between matched pairs. This, in turn, translates into reduced sensitivity to biases due to hidden covariates under certain models of analysis (see chapter 4 of [21]).

The previous findings should not be surprising because cardinality matching is designed to explicitly optimize sample size and directly constrain covariate balance, and optimal subset matching is designed to minimize the covariate distances between matched pairs given a threshold distance. On the other hand, nearest neighbor matching is a greedy method with no optimality guarantees regarding any of the three previous comparison criteria (covariate balance, sample size, and covariate distances).

In terms of treatment effect estimates, our simulation study results suggest that, with a simple differences in means estimator, better covariate balance translates into better estimates. In general, the estimates obtained with cardinality matching have lower RMSEs, except for the case in which it only requires the means to have absolute standardized differences smaller than 0.1. In particular, when the outcome is known to be exactly a linear combination of the observed covariates, tight mean balance appears to be enough to remove systematic biases. However, in practice the investigator does not really know this, and the covariates may affect the outcome in a nonlinear way, so it is preferable to match with fine balance for all the covariates in addition to a tight mean balance constraint. The inclusion of strength-2 balancing constraints for all covariates could significantly improve the estimation when there are important interaction terms affecting the outcome. However, the number of these type of restrictions grows quickly with each additional covariate, and this could be very demanding for some datasets. Thus, it is advised that if the researcher conjectures a possible interaction term between some specific covariates, strength-k balancing for those covariates should be performed if feasible. Note that even when strength-k balancing is not used, fine balance with mean balance constraints provides the best results when estimating a treatment effect under all the scenarios and outcomes considered.

A last but important point observed in this simulation study is the relatively poor performance in every aspect evaluated of the matching that only requires the standardized differences in means of all covariates to be below 0.1. This suggests that the common rule of thumb of balancing covariates so that their absolute standardized differences in means are not greater than 0.1 is typically not enough, and that stronger forms of balance should be pursued in practice when using a simple difference in means effect estimator.

References

1. Cochran W, Rubin D. Controlling bias in observational studies: A review. *Sankhyā: The Indian Journal of Statistics, Series A* 1973; :417–446.
2. Cochran WG, Moses LE, Mosteller F. *Planning and analysis of observational studies*. Wiley: New York, 1983.
3. Stuart EA. Matching methods for causal inference: A review and a look forward. *Statistical Science* 2010; **25**(1):1–21.
4. Stuart EA, Rubin DB. *Matching Methods for Causal Inference*, chap. 11. Thousand Oaks, CA: Sage Publications, 2007; 155–176.
5. Snedecor GW, Cochran WG. *Statistical Methods*. Iowa State University Press, 1980.
6. Rosenbaum PR. Heterogeneity and causality: Unit heterogeneity and design sensitivity in observational studies. *The American Statistician* April 2005; **59**(2):147–152.
7. Dehejia R, Wahba S. Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. *Journal of the American Statistical Association* 1999; **94**(443):1053–1062.
8. Rosenbaum PR. Optimal matching of an optimally chosen subset in observational studies. *Journal of Computational and Graphical Statistics* 2012; **21**(1):57–71.
9. Zubizarreta JR, Paredes RD, Rosenbaum PR. Matching for balance, pairing for heterogeneity in an observational study of the effectiveness of for-profit and not-for-profit high schools in chile. *Annals of Applied Statistics* 2014; **8**(1):204–231.
10. Zubizarreta JR. Using mixed integer programming for matching in an observational study of kidney failure after surgery. *Journal of the American Statistical Association* 2012; **107**(500):1360–1371.
11. Rosenbaum PR, Ross RN, Silber JH. Minimum distance matched sampling with fine balance in an observational study of treatment for ovarian cancer. *Journal of the American Statistical Association* 2007; **102**(477):75–83.
12. Hsu JY, Zubizarreta JR, Small DS, Rosenbaum PR. Strong control of the family-wise error rate in observational studies that discover effect modification by exploratory methods. *Biometrika* 2015; **102**(4):767–782.
13. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika* 1983; **70**(1):41–55.
14. Gu XS, Rosenbaum PR. Comparison of multivariate matching methods: Structures, distances, and algorithms. *Journal of Computational and Graphical Statistics* 1993; **2**(4):405–420.
15. Dehejia R, Wahba S. Propensity score-matching methods for nonexperimental causal studies. *The Review of Economics and Statistics* 2002; **84**(1):151–161.
16. Austin PC. Some methods of propensity-score matching had superior performance to others: Results of an empirical investigation and monte carlo simulations. *Biometrical Journal* 2009; **51**(1):171–184.
17. Austin PC. Optimal caliper widths for propensity-score matching when estimating differences in means and differences in proportions in observational studies. *Pharmaceutical Statistics* 2011; **10**(2):150–161.

18. Austin PC. A comparison of 12 algorithms for matching on the propensity score. *Statistics in Medicine* 2014; **33**(6):1057–1069.
19. Rubin DB. Matching to remove bias in observational studies. *Biometrics* 1973; **29**:159–183.
20. Rosenbaum PR, Rubin DB. Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician* 1985; **39**(1):33–38.
21. Rosenbaum PR. *Observational Studies*. Springer, 2002.
22. Rosenbaum PR. Optimal matching for observational studies. *Journal of the American Statistical Association* 1989; **84**:1024–1032.
23. Burkard R, Dell'Amico M, Martello S. *Assignment Problems*. Society for Industrial and Applied Mathematics: Philadelphia, PA, USA, 2009.
24. Ahuja R, Magnanti T, Orlin J. *Network Flows: Theory, Algorithms and Applications*. Prentice Hall: New Jersey, 1993.
25. Bertsimas D, Tsitsiklis JN. *Introduction to Linear Optimization*. Athena Scientific, Dynamic Ideas, 1997.
26. Kuhn H. The Hungarian Method for the Assignment Problem. *Naval research logistics quarterly* 1955; **2**(1-2):83–97.
27. Bertsekas DP. A new algorithm for the assignment problem. *Mathematical Programming* 1981; **21**:152–171.
28. Papadimitriou C. *Computational Complexity*. Addison-Wesley: Reading (Mass.), 1994.
29. Hansen BB, Klopfer SO. Optimal full matching and related designs via network flows. *Journal of Computational and Graphical Statistics* 2006; **15**(3):609–627.
30. Rosenbaum PR. *Design of Observational Studies*. Springer, 2010.
31. Baiocchi M. Designing robust studies using propensity score and prognostic score matching. *Chapter 3 in Methodologies for Observational Studies of Health Care Policy, Dissertation, Department of Statistics, The Wharton School, University of Pennsylvania* 2011; .
32. Hansen BB. Flexible, optimal matching for observational studies. *R News* 2007; **7**:18–24.
33. Zubizarreta JR, Kilcioglu C. `designmatch`*: Construction of Optimally Matched Samples for Randomized Experiments and Observational Studies that are Balanced by Design* 2016. `R` package version 0.2.0.
34. Yang F, Zubizarreta JR, Small DS, Lorch S, Rosenbaum PR. Dissonant conclusions when testing the validity of an instrumental variable. *The American Statistician* 2014; **68**(4):253–263.
35. Keele L, Titiunik R, Zubizarreta JR. Enhancing a geographic regression discontinuity design through matching to estimate the effect of ballot initiatives on voter turnout. *Journal of the Royal Statistical Society: Series A* 2015; **178**:223–239.
36. Normand SLT, Landrum MB, Guadagnoli E, Ayanian JZ, Ryan TJ, Cleary PD, McNeil BJ. Validating recommendations for coronary angiography following acute myocardial infarction in the elderly: A matched analysis using propensity scores. *Journal of Clinical Epidemiology* 2001; **54**(4):387–398.
37. Zubizarreta JR. Stable weights that balance covariates for estimation with incomplete outcome data. *Journal of the American Statistical Association* 2015; **110**(511):910–922.
38. Rosenbaum PR. An exact distribution-free test comparing two multivariate distributions based on adjacency. *Journal of the Royal Statistical Society: Series B* 2005; **67**(4):515–530.
39. Imai K, King G, Stuart E. Misunderstandings among experimentalists and observationalists about causal inference. *Journal of the Royal Statistical Society, Series A* 2008; **171**(2):1–22.
40. Hansen BB, Bowers J. Covariate balance in simple, stratified and clustered comparative studies. *Statistical Science* 2008; **23**(2):219–236.
41. Iacus SM, King GK, Porro G. Causal inference without balance checking: Coarsened exact matching. *Political Analysis* 2012; **20**(1):1–24.
42. Franklin JM, Rassen JA, Ackermann D, Bartels DB, Schneeweiss S. Metrics for covariate balance in cohort studies of causal effects. *Statistics in Medicine* 2014; **33**(10):1685–1699.
43. Heller R, Rosenbaum PR, Small DS. Using the cross-match test to appraise covariate balance in matched pairs. *The American Statistician* 2010; **64**(4):299–309.
44. Haviland A, Nagin D, Rosenbaum P. Combining propensity score matching and group-based trajectory analysis in an observational study. *Psychological Methods* 2007; **12**(3):247.
45. Heller R, Rosenbaum PR, Small DS. Split samples and design sensitivity in observational studies. *Journal of the American Statistical Association* 2009; **104**(487):1090–1101.
46. Zhang K, Small DS, Lorch S, Srinivas S, Rosenbaum PR. Using split samples and evidence factors in an observational study of neonatal outcomes. *Journal of the American Statistical Association* 2011; **106**(494):511–524.
47. Rosenbaum PR, Rubin DB. The bias due to incomplete matching. *Biometrics* 1985; **41**(1):103–116.
48. Hill JL. Discussion of research using propensity-score matching: Comments on a critical appraisal of propensity-score matching in the medical literature between 1996 and 2003 by peter austin, statistics in medicine. *Statistics in Medicine* 2008; **27**(12):2055–2061, doi:10.1002/sim.3245. URL `http://dx.doi.org/10.1002/sim.3245`.
49. Traskin M, Small D. Defining the study population for an observational study to ensure sufficient overlap: a tree approach. *Statistics in Biosciences* 2011; **3**:94–118.
50. Fogarty C, Mikkelsen M, Gaieski D, Small D. Discrete optimization for interpretable study populations and randomization inference in an observational study of severe sepsis mortality. *Journal of the American Statistical Association* 2016; :forthcoming.

51. Lee BK, Lessler J, Stuart EA. Improving propensity score weighting using machine learning. *Statistics in Medicine* 2010; **29**(3):337–346, doi: 10.1002/sim.3782. URL http://dx.doi.org/10.1002/sim.3782.

52. Imbens GW, Rubin DB. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge University Press, 2015.

53. Crump RK, Hotz VJ, Imbens GW, Mitnik OA. Dealing with limited overlap in estimation of average treatment effects. *Biometrika* 2009; **96**(1):187–199.

54. Imbens GW. Matching methods in practice: Three examples. *Journal of Human Resources* 2015; **50**(2):373–419.

55. King G, Lucas C, Nielsen R. The balance-sample size frontier in matching methods for causal inference 2015.

56. Rubin DB. For objective causal inference, design trumps analysis. *The Annals of Applied Statistics* 2008; **2**(3):808–840. URL http://www.jstor.org/stable/30245110.