Journal of Evaluation in Clinical Practice
International Journal of Public Health Policy and Health Services Research

Journal of Evaluation in Clinical Practice ISSN 1365-2753

Using balance statistics to determine the optimal number of controls in matching studies

Ariel Linden DrPH[1,2] and Steven J. Samuels PhD[3]

[1] President, Linden Consulting Group, Ann Arbor, Michigan, USA
[2] Adjunct Associate Professor, Department of Health Policy & Management, School of Public Health, University of Michigan, Ann Arbor, Michigan, USA
[3] Adjunct Associate Professor, Department of Epidemiology & Biostatistics, School of Public Health, State University of New York, Albany, New York, USA

Keywords
balance, bias, matching, observational studies, propensity score

Correspondence
Dr Ariel Linden
Linden Consulting Group, LLC
1301 North Bay Drive
Ann Arbor
MI 48103
USA
E-mail: alinden@lindenconsulting.org

Accepted for publication: 2 July 2013

doi:10.1111/jep.12072

Financial Support: None.

Authors' Note: All of the proposed methods in this paper can be implemented via companion software, for Stata, which is available upon request from the authors.

Abstract

When a randomized controlled trial is not feasible, investigators typically turn to matching techniques as an alternative approach to evaluate the effectiveness of health care interventions. Matching studies are designed to minimize imbalances on measured pre-intervention characteristics, thereby reducing bias in estimates of treatment effects. Generally, a matching ratio up to 4:1 (control to treatment) elicits the lowest bias. However, when matching techniques are used in prospective studies, investigators try to maximize the number of controls matched to each treated individual to increase the likelihood that a sufficient sample size will remain after attrition. In this paper, we describe a systematic approach to managing the trade-off between minimizing bias and maximizing matched sample size. Our approach includes the following three steps: (1) run the desired matching algorithm, starting with 1:1 (one control to one treated individual) matching and iterating until the maximum desired number of potential controls per treated subject is reached; (2) for each iteration, test for covariate balance; and (3) generate numeric summaries and graphical plots of the balance statistics across all iterations in order to determine the optimal solution. We demonstrate the implementation of this approach with data from a medical home pilot programme and with a simulation study of populations of 100 000 in which 1000 individuals receive the intervention. We advocate undertaking this methodical approach in matching studies to ensure that the optimal matching solution is identified. Doing so will raise the overall quality of the literature and increase the likelihood of identifying effective interventions.

Introduction

In health care settings, conducting a randomized controlled trial (RCT) to evaluate the effectiveness of programmes and other interventions is often infeasible due to logistical, practical or ethical reasons. When only observational data are available, investigators use matching techniques to create a control group that is similar to the treatment group. Matching is performed on observed pre-intervention characteristics only, and unlike RCTs, must assume that the unknown characteristics will not bias the results [1].

Generally, a matching ratio of up to 4:1 (controls to treated subjects) elicits the lowest bias in treatment effect estimates [2–5]. Higher ratios typically increase bias because each additional matched control will be less comparable to the treated subject than the first matched control, and fewer controls will be available overall for matching to treated subjects later in the matching process. In prospective studies where attrition is likely, there is an additional consideration; sufficient controls must be matched to each treated subject at the outset to ensure that at least one control will remain matched to each treated individual at the study's conclusion. If no acceptable control matches are found, unmatched treated individuals will be dropped from the analysis [6]. Thus, a challenging issue facing investigators who use a matching strategy is to balance the trade-off between reducing bias and maximizing the matched sample size.

The purpose of the current paper is to describe a systematic approach to choosing the optimal number of controls per case in a matching study in which the investigator has the ability to choose more than one control for every treated subject. First, the investigator specifies the maximum number of controls per treated subject being considered. Next, for each potential number of controls, from 1 to *k*, a matching algorithm creates a matched data set with up to the specified number of controls per case and generates balance statistics [standardized differences in means, variance ratios (VRs)] for each variable used for matching. Finally, the balance statistics are summarized and plotted against the number of controls per treated individual, allowing the investigator to visually identify the number of controls that provides the best trade-off between bias and sample size. While we utilize a propensity score approach, a strength of this methodology is that it can be implemented in conjunction with any matching procedure that offers a *k*:1 solution (we refer readers to Stuart [7] for a comprehensive treatment of available matching approaches and procedures, and Caliendo & Kopeinig [8] for a more tailored discussion on propensity score matching).

The paper is organized as follows: In the next section, ‘Measures of Covariate Balance’, we describe the two numeric measures of covariate balance used in our approach. We also explain how these measures, which are commonly used to assess covariate balance in 1:1 matching, are modified for *k*:1 matching. In the subsequent section, ‘Example 1: A Medical Home Pilot Programme’, we demonstrate our approach using data from a primary-care based medical home pilot programme. In the following section titled ‘Example 2: A Monte Carlo Simulation Study’, we demonstrate the approach on simulated populations of 100 000 people of whom 1000 are exposed to a hypothetical prospective intervention. In the ‘Discussion’, we summarize our findings and offer recommendations for researchers employing these techniques. We close with a set of concluding remarks.

Measures of covariate balance

In order to ensure valid results, an intervention study, whether randomized or observational, must have treatment and control groups that are comparable on pre-intervention characteristics. While the nature of randomization should produce balance on both observed and unobserved covariates, in observational studies, we cannot make this assumption and therefore assess covariate balance on observed characteristics alone.

Generally, covariate balance is assessed by both graphical and numerical diagnostics [7,9]. Graphic displays such as box plots and density probability plots [10,11] provide a visual snapshot of balance across the distribution of a covariate. Although these plots are useful, their interpretation is subjective and they generally do not provide quantifiable summaries that can be used for analytic comparisons. Numerical measures, on the other hand, allow for objective criteria to be used to determine covariate balance and to generate comparisons across different matching solutions. Therefore, in the current study, we rely on numerical balance measures, and we focus on two commonly used measures: the standardized difference between treatment and control means, and the ratio of treatment and control variances (‘the variance ratio’).

Standardized mean difference (SMD)

With no matching or 1:1 matching, the SMD for a given covariate *j* is defined as [12]:

$$smd_j = \frac{|\bar{X}_{jT} - \bar{X}_{jC}|}{\sqrt{\frac{(S_{jT})^2 + (S_{jC})^2}{2}}} \tag{1}$$

where the numerator is the absolute difference in means between the treatment and control groups (denoted as *T* and *C*, respectively) and the denominator is a 50:50 pooled standard deviation. Dichotomous covariates can also be tested for balance using this equation or using a formula specific to proportions [9]. This measure is dimensionless and is not sensitive to sample size. While there is no empirical evidence to support the use of any particular cut-off point to define imbalance, Normand *et al.* [13] suggest that a standardized difference greater than 0.10 is indicative of imbalance and Rubin [14] suggests a cut-off of 0.25. Alternatively, since the standardized difference is a version of Cohen’s *d* statistic for effect size [15], one could also argue for a cut-off of 0.20, which Cohen termed a ‘small’ effect. Thus, the investigator has a rather broad range of acceptable cut-offs to choose from.

When multiple covariates are tested for balance, an overall summary statistic of the individual SMDs is helpful. This can take many forms, such as the average, the median or the percentage of covariates with standardized differences less than a designated limit. We use the average SMD, defined for *J* variables as follows:

$$SMD = \frac{1}{J}\sum_{j=1}^{J} smd_j \tag{2}$$

which is the sum of the individual absolute standardized differences divided by the number of covariates assessed. When comparing different matching solutions, higher values of SMD indicate greater imbalance across covariates.

VRs

As balance is not only a property of the sample means of a covariate but of the overall distribution, higher-order sample moments of the distribution should be evaluated as well. Rubin [14] proposes the use of the ratio of treated and control variances as a balance measure. In the case of 1:1 matching for continuous variables, the VR for a given covariate *j* is as follows:

$$VR_j = \frac{(S_{j,T})^2}{(S_{j,C})^2} \tag{3}$$

where the numerator is the variance of the covariate in the treated group and the denominator is the variance of the covariate in the control group. Better balance is defined by values close to 1.0. Rubin [14] suggests that variables are out of balance if the VR is greater than 2.0 or less than 0.5. As this ‘two-sided’ definition of imbalance leads to difficulties when a summary VR must be created, we consider a modified version. Recall that the SMD is always positive, regardless of which group’s mean value is larger. We define an analogous variance ratio, *VR**, which is always greater than 1, regardless of which group’s variance is larger, with a value greater than 2.0 indicating imbalance:

$$VR^*j = \frac{(S_{j\max})^2}{(S_{j\min})^2} \tag{4}$$

Table 1 Baseline (12 months) characteristics of programme participants (treated) and non-participants (non-treated) (from [18])

Variable*	Treated (n = 374)	Non-treated (n = 1628)	Standardized difference†	Variance ratio‡
Demographic characteristics				
Age	54.87 (6.71)	43.44 (11.99)	1.177	3.192
Female (%)	56.4	49.6	0.137	
Utilization and cost				
Primary care visits	11.29 (7.30)	4.63 (4.35)	1.111	2.820
Other outpatient visits	18.03 (16.65)	7.25 (10.61)	0.772	2.463
Laboratory tests	6.09 (5.27)	2.38 (3.31)	0.844	2.542
Radiology tests	3.20 (4.46)	1.31 (2.48)	0.524	3.225
Prescriptions filled	40.59 (29.96)	11.95 (17.14)	1.174	3.055
Hospitalizations	0.24 (0.52)	0.07 (0.29)	0.403	3.239
Emergency department visits	0.39 (1.03)	0.16 (0.50)	0.287	4.232
Home-health visits	0.09 (0.88)	0.01 (0.38)	0.108	5.462
Total costs	8237 (9830)	3047 (5817)	0.643	2.856
Average§			0.653	2.897

Notes:

*Unless otherwise noted, variables presented are means and standard deviations.

†Standardized differences were calculated using Equation 1 for all variables.

‡Variance ratios are calculated using Equation 4 for continuous variable only.

§The averages are derived using Equations 2 and 5, respectively.

where the maximum of the treated and untreated squared standard deviations is in the numerator and the minimum is in the denominator.

To summarize the VRs of multiple covariates, we take the geometric mean of the *VR** values and refer to it as the *Geometric Mean Variance Ratio* (*GMVR*):

$$GMVR = \left(\prod_{j=1}^{J} VR_j^* \right)^{1/J} \qquad (5)$$

As with the individual *VR** values, the *GMVR* is always greater than 1 and, as before, values greater than 2.0 reflect imbalance.

Numeric diagnostics under a *k*:1 matching strategy

For *k*:1 matching, one can consider either (1) *fixed ratio matching,* in which exactly *k* controls are matched to each treated subject, or (2) *variable ratio matching,* in which up to *k* controls are matched to each treated subject with the potential for treated subjects to be assigned a varying number of controls. For example, with 3:1 variable ratio matching, one subject may only have one matched control while another subject may have three controls. Variable ratio matching has been shown to have better bias reduction properties than fixed ratio matching [16], and so we follow a variable ratio matching approach in the current study.

When variable ratio matching is employed, weights must be incorporated into the analysis to avoid bias in estimated treatment effects [17]. When one control is matched to one treated subject, the mean covariate value for that control is the covariate value itself. However, when two controls are matched to one treated subject, the mean covariate value is an average of the two controls' values. For most matching studies, the goal is to estimate the average treatment effect in the treated (ATT), and the weight of the treated subject must equal the sum of the weights for its matched controls. While any weighting system meeting this criterion will do, we give each treated subject a weight of 1, and each of its controls a weight of 1/*k* [17]. For example, if three controls are matched to one treated subject, each control gets a weight of 1/3. When utilizing variable ratio matching, weighted means and standard deviations replace their unweighted counterparts in the formulas for SMDs and VRs (equations 1–5).

Example 1: A medical home pilot programme

Data

We use data from a primary care-based medical home pilot programme that invited patients to enrol if they had a chronic illness or were predicted to have high costs in the following year. The goal of the programme was to lower health care costs for programme participants by providing intensified primary care (see Linden [18], for a more comprehensive description). The retrospectively collected data consist of observations for 374 programme participants and 1628 non-participants. Table 1 describes the pre-intervention characteristics of the treated and non-treated groups, together with their unadjusted standardized differences and VRs. As shown, the treated group differed markedly from the non-treated group on every variable. On average, treated individuals were older, were less likely to be female, and had higher utilization and costs than non-treated individuals. All standardized differences at the individual variable level were greater than 0.10 and all of the VRs were greater than 2.0. Both the summary average SMD and geometric mean VR exceeded those cut-points as well (SMD = 0.653, GMVR = 2.897). These significant imbalances highlight the need for an effective matching strategy to create a control group that more closely resembles the treatment group.

Table 2 Weighted sample means* by *k*:1 matching solutions (up to four controls per treated subject)

	Maximum number of controls per treated							
	1		2		3		4	
Variable	Treated	Control	Treated	Control	Treated	Control	Treated	Control
n	271	271	225	403	206	496	196	553
Age	53.92	54.75	53.38	54.52	53.10	54.39	52.96	54.35
Female (%)	55.72	55.72	53.33	52.67	54.37	53.00	56.12	53.44
Primary care visits	9.73	9.43	9.32	8.83	9.34	8.66	9.48	8.65
Other outpatient visits	15.26	14.58	14.84	13.82	14.91	13.62	14.96	13.71
Laboratory tests	5.10	4.90	4.64	4.51	4.68	4.47	4.76	4.51
Radiology tests	2.99	2.75	3.00	2.46	3.10	2.41	3.17	2.39
Prescriptions filled	32.94	32.24	31.27	30.02	30.87	29.67	30.84	30.19
Hospitalizations	0.17	0.16	0.16	0.13	0.16	0.13	0.16	0.14
Emergency department visits	0.34	0.33	0.32	0.31	0.35	0.29	0.36	0.28
Home-health visits	0.03	0.09	0.02	0.06	0.02	0.06	0.03	0.05
Total costs	6613	6364	6336	5685	6325	5903	6468	5892

Notes: *Except for *n* which is reported as a count, and female, which is reported as a percentage.

Methods

Although an array of matching strategies could have been employed, we selected propensity score matching, due to its widespread use in the medical literature [19–21]. The propensity score reflects the probability of assignment to the treatment group conditional on observed covariates [22]. It reduces bias by controlling for pre-intervention differences between treated and non-treated groups. Propensity scores are generally derived from a logistic regression equation that reduces each participant's set of covariates to a single score. Conditional on a well-constructed propensity score, pre-treatment covariates will be independent of group assignment and will be distributed similarly in both groups. When correlation of covariates and treatment assignment is removed, the covariates will not confound estimated treatment effects [22].

In our example, the propensity score was estimated from a logit model in which the treatment variable was regressed on the 11 baseline covariates listed in Table 1, entered as main effects. Then, to select the controls on which to calculate the imbalance measures described above, we performed variable ratio *k*:1 matching with a modified version of FGMATCH, a user-written Stata command for greedy matching [23], which in turn, is an enhanced version of a similar programme written for SAS by Parsons [24]. In FGMATCH, one specifies the number *k* for variable ratio *k*:1 matching. Before the process begins, controls are randomly ordered. Matching is performed without replacement, so that controls are matched to at most one case. The criterion for matching a treated subject to a given control is the number of decimal digits to which their propensity scores agree. In our application, an attempt was first made to match on five decimal digits; if fewer than *k* controls were matched to a case, matching on four decimal digits was attempted, continuing to a match on a single digit.

To compute the imbalance measures, we modified PBALCHK [25], another Stata command contributed by the author of FGMATCH, to estimate the unweighted and weighted SMD and VR statistics as described earlier. We report results for *k* ranging from one to four and also show the number of treated and controls that were matched for each value of *k*. All analyses were conducted using Stata 12.1 (StataCorp., College Station, TX, USA).

Results

Table 2 presents sample sizes and weighted covariate means for treated and matched controls. Note that as the maximum number of controls per case increases, the treated and untreated means remain relatively unchanged, but the total number of matched treated subjects decreases.

Table 3 presents the absolute standardized differences in weighted covariate means for each of the variable ratio matching solutions ranging from one up to four controls per treated individual (unmatched values for standardized differences and VRs are taken from Table 1 and presented in Tables 3 and 4 for convenience). The average standardized difference remains below 0.10 for all matching scenarios, indicating that, on average, all solutions provide good covariate balance. However, the number of individual covariates with SMD values greater than 0.10 increases from one (for the 1:1 solution) up to four (for the 3:1 and 4:1 solutions). No covariates would be considered imbalanced if the threshold was set at either 0.20 [15] or 0.25 [14].

Table 4 presents the weighted VRs for each of the matching solutions ranging from one up to four controls per treated individual. As shown, the geometric mean VR is within the range of 1.0 and 2.0 for all matching scenarios, indicating once again that on average, all solutions provide good covariate balance. At the individual covariate level, however, laboratory testing rates and home-health visit rates persistently fall outside of this range, while radiology testing falls outside of the range at 4:1 matching. These results are explained by consistently higher variances in the control group for these variables which, in turn, is driven by the much larger ranges of values in these variables among controls compared to those in the treated group.

Figure 1 provides a graphical display of the average standardized difference and GMVR for each *k*:1 solution. While this figure

Table 3 Standardized differences* for unmatched and *k*:1 matching solutions (up to four controls per treated subject)

		Maximum number of controls per treated			
Variable	Unmatched	1	2	3	4
Age	1.177	0.082	0.113	0.134	0.137
Female	0.137	0.030	0.004	0.015	0.061
Primary care visits	1.111	0.072	0.099	0.128	0.135
Other outpatient visits	0.772	0.040	0.101	0.118	0.096
Laboratory tests	0.844	0.080	0.023	0.048	0.060
Radiology tests	0.524	0.066	0.127	0.174	0.183
Prescriptions filled	1.174	0.003	0.057	0.062	0.043
Hospitalizations	0.403	0.049	0.054	0.065	0.065
Emergency department visits	0.287	0.029	0.031	0.083	0.094
Home-health visits	0.108	0.145	0.112	0.085	0.069
Total costs	0.643	0.033	0.083	0.090	0.075
Average standardized difference†	0.653	0.057	0.073	0.091	0.093

Notes:
*Standardized differences were calculated using Equation 1 with and without weights for the unmatched sample and for all *k*:1 solutions, accordingly.
†Average standardized difference was derived using Equation 2.

Table 4 Variance ratios* for unmatched and *k*:1 matching solutions (up to four controls per treated subject). Values presented are for continuous variables only

		Maximum number of controls per treated			
Variable	Unmatched	1	2	3	4
Age	3.192	1.186	1.175	1.274	1.290
Primary care visits	2.820	1.143	1.062	1.010	1.023
Other outpatient visits	2.463	1.332	1.272	1.204	1.185
Laboratory tests	2.542	1.936	2.306	2.259	2.247
Radiology tests	3.225	1.150	1.586	1.872	2.043
Prescriptions filled	3.055	1.373	1.386	1.414	1.412
Hospitalizations	3.239	1.050	1.006	1.044	1.085
Emergency department visits	4.232	1.580	1.703	1.906	1.939
Home-health visits	5.462	5.648	4.937	3.609	2.945
Total costs	2.856	1.251	1.042	1.017	1.188
Geometric mean variance ratio†	3.309	1.462	1.473	1.467	1.481

Notes:
*Variance ratios were calculated using Equation 4 with and without weights for the unmatched sample and for all *k*:1 solutions, accordingly.
†Geometric mean variance ratio was derived using Equation 4.

provides the same information as presented in Tables 3 and 4, the graphic form allows the investigator to examine multiple aspects of the data simultaneously.

Example 2: A Monte Carlo simulation study

In this section, we use Monte Carlo simulation to demonstrate the performance of our iterative approach in identifying the optimal number of controls per treated individual, when the pool of eligible controls is extremely large relative to the number of treated individuals. This is a common scenario for many interventions implemented in health plans, such as disease management or case management programmes. These populations typically experience a high turnover rate, requiring programme evaluators to maximize the number of controls matched to each treated individual at the outset, in an effort to maintain sufficient sample size at the end of the study. With simulated data, we have the additional ability to generate an 'actual' treatment effect and then test which matching approach provides the most accurate solution (i.e. derives the closest treatment effect to the actual).

Design

We began by generating a pseudo-population of 100 000 individuals comprised of 1000 treated and 99 000 untreated individuals. Following the data generating process described by Austin [26], we created five continuous and five dichotomous covariates. For the continuous covariates, untreated individuals received randomly generated values from a normal distribution with mean of 0 and standard deviation of 1. Treated individuals received randomly generated values from a normal distribution with means of 0.20, 0.30, 0.40, 0.50 and 0.60, and a standard deviation of 1, for each of the five continuous covariates respectively. For the dichotomous

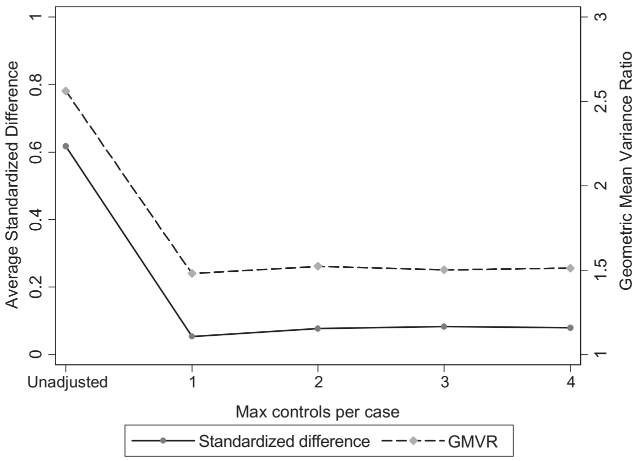

Figure 1 Average standardized difference and geometric mean variance ratio for each *k*:1 solution (up to four controls per treated subject).

covariates, untreated individuals received uniformly distributed random variates to simulate prevalence rates of 0.10 through 0.50 for the five dichotomous covariates, respectively. Treated individuals received uniformly distributed random variates to simulate prevalence rates of 0.168, 0.331, 0.492, 0.642 and 0.776 for the five dichotomous covariates, respectively. This setup produced standardized differences of 0.20, 0.30, 0.40, 0.50 and 0.60 between treated and the unmatched pool of untreated individuals for both the continuous and dichotomous covariates, respectively. Next, we randomly generated a continuous outcome for each individual:

$$Y_{\mathrm{i}} = -2.8 + 1.5C_{1\mathrm{i}} + 2C_{2\mathrm{i}} + 3C_{3\mathrm{i}} + 4C_{4\mathrm{i}} + 5C_{5\mathrm{i}} + 5B_{1\mathrm{i}} + 4B_{2\mathrm{i}} \\ + 3B_{3\mathrm{i}} + 2B_{4\mathrm{i}} + 1.5B_{5\mathrm{i}} + T_{\mathrm{i}} + \varepsilon_{\mathrm{i}}$$

where Y_{i} is a function of the treatment variable T (with a treatment effect of 1.0), the five continuous covariates (C_1-C_5), the five binary covariates (B_1-B_5), and a normally distributed error term ε with mean 0 and standard deviation 2 [26].

For each individual in the pseudo-population, the propensity score was estimated from a logit model in which the treatment variable was regressed on the 10 covariates, entered as main effects. Following the procedure described in Example 1, we performed iterative *k*:1 matching up a maximum of 70 matched controls for each treated individual. Because of the large number of controls, we required matching on the propensity scores to the nearest two decimal digits, not one as in the first example. At each iteration, we calculated the standardized difference for each of the 10 covariates using Equation 1 and the average SMD, using Equation 2. The treatment effect was estimated by regressing the outcome on the treatment variable using the weight generated to account for *k*:1 matching as described earlier. This simulation was then repeated 100 times and the results were averaged across all the simulated data sets.

Results

The online Supporting Information presents tables with the simulation results averaged for each of the 70 matched scenarios. Supporting Information Table S1 presents the sample sizes, standardized differences for each of the 10 covariates, the actual number of imbalanced covariates (i.e. standardized difference >0.10) and the average standardized difference across all 10 covariates. We omit VRs from this analysis because, with exception of one or two scenarios, all were very close to 1.0. By design, all covariates in the unmatched population were imbalanced, producing an average SMD of approximately 0.40. Not until a matching ratio of up to 61 controls to each treated subject did the average SMD again cross over the 0.10 threshold. The actual number of covariates out of balance remained at zero until a matching ratio of up to 47 controls to a treated subject, and then rose to a maximum of six for the matching ratio of up to 70 controls per treated subject. The number of treated individuals dropped due to lack of good matches was relatively small overall, reaching a maximum of 48 in the highest matching ratio solution.

Supporting Information Table S2 presents the estimated treatment effect, absolute bias, mean squared error (MSE) and root mean squared error (RMSE) for each of the 70 matching ratios. As shown, overall matching ratios between one and four controls per treated subject elicited the lowest bias in treatment effect estimates, with the 1:1 ratio having the lowest absolute bias (e.g. the mean estimated treatment effect of 1.007 minus the true treatment effect of 1.0 minus = 0.007, or 0.7%), and the 2:1 ratio producing the lowest MSE and RMSE. Beyond the 4:1 matching ratio, these measures of bias increase substantially. There is also no gain in study precision when matching more than four controls to each treated individual, as standard errors for the estimated treatment effect barely change for $k > 4$.

Discussion

Matching is often favoured over other adjustment techniques, such as regression analysis, because it decouples the design from the analysis of treatment effects [14]. A matching strategy, however, requires that the investigator balance baseline covariates to reduce biased treatment effect estimates, and in a longitudinal study, to ensure sufficient matched-sets are available for the final analysis. Although the most common propensity-score matching technique found in the medical literature is 1:1 matching [19,20], rarely do authors describe an assessment of whether this matching solution is optimal. It is therefore likely that 1:1 matching is selected *a priori* as the method of choice, which misses the potential to improve the evaluation. In this paper, we develop an approach that can be used by investigators to identify the optimal number of controls. Our approach relies on standardized differences and VRs as numeric measures of covariate balance. However, our approach can be adapted to any numeric balance diagnostic that accommodates weights, such as the standardized difference in variances [27] or the multivariate imbalance measure based on the L1 distance [28].

In our medical home pilot programme example, the two balance measures suggested that overall covariate balance is achieved in variable ratio matching of up to four controls per treated individual; however, some individual covariates remain imbalanced. Investigators may find it important to consider which covariates are imbalanced for a given matching strategy. For example, Table 3 indicates that age and radiology test rates are imbalanced (standardized differences > 0.10) at all matching levels greater than 1:1, while home health visit rates are imbalanced in the 1:1 and 2:1 solutions. Table 4 shows that laboratory test rates and

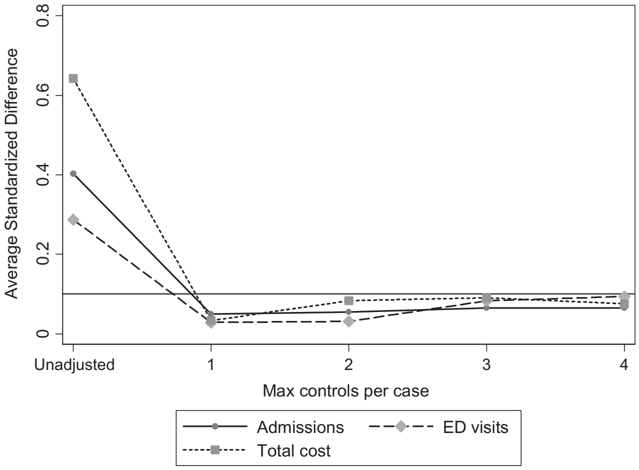

Figure 2 Standardized difference for individual covariates (emergency department visits, admissions and total costs) at each *k*:1 solution (up to four controls per treated subject). Horizontal line at 0.10 represents an upper limit of balance for the standardized difference.

home health visit rates have imbalanced VRs. Both theoretical and applied expertise is necessary to determine which imbalances are acceptable and which are not. Here, one could argue that the most important covariates requiring balance are emergency department visits, hospitalizations and costs, as these are the outcomes expected to be directly impacted by the intervention [29]. Visual displays of the balance statistics for these individual covariates can assist the investigator in quickly examining multiple aspects of the variables simultaneously. As illustrated in Fig. 2, the standardized differences for emergency department visits, hospitalizations and costs are within the acceptable limit of 0.10 across all matching solutions. Rather than discarding imbalanced covariates, residual imbalances can be further adjusted within a regression framework during the analysis stage [30]. For large imbalances, however, regression adjustment would consist of an extrapolation between treated and controls whose distributions on the covariate do not overlap. If these imbalances occur on essential covariates, the investigator may conclude that these data do not permit credible estimation of the intervention effect.

Another important factor to consider when choosing a matching solution is the number of treated observations that are lost as *k* increases and the pool of controls is diminished. Investigators may feel that generalizability is compromised with a sizeable loss of treated units. In practice, the choice of *k* will be based on the particular characteristics of the study. Specifically, in a retrospective study where all the available data have been collected, the investigator should choose the matching solution that retains the largest number of treated individuals (which will most likely be the 1:1 solution). In the context of the medical home pilot data, about 28% (103 of 374) of treated subjects could not be matched to even one control, and attempts to match to more than one control resulted further losses of treated subjects for whom matches could not be found (Table 2). When retaining the maximum number of treated individuals is critical, a possible solution to consider is implementing a matching 'with replacement' strategy. This procedure involves replacing a matched control back into the dataset so that it may be matched to additional treated individuals. One possible drawback is that in some situations, certain controls may be used an inordinate number of times, thereby exerting excessive influence on the treatment effect estimate. Thus, when using matching with replacement, the number of times each control is matched should be monitored [7].

Our second example simulates a prospective study in which high rates of attrition are anticipated over time. In this simulation, we see that no covariates are imbalanced up to a ratio of 47 controls per treated subject, and the average SMD does not surpass the 0.10 threshold until a ratio of 61 controls per treated subject is reached. These results underscore the danger of blind dependence on 'standard' guidelines for defining balance. Although all covariates remain 'balanced' up to $k = 47$ according to the 0.10 criterion, the absolute bias in the treatment effect and RMSE is approximately 386% at that matching solution (Supporting Information Table S2). In fact, our simulation study confirmed findings from other research suggesting that matching ratios of up to 4:1 elicit the lowest bias in treatment effect estimates [2–5]. The absolute bias in the estimated treatment effect nearly doubles from 8.5% to 16.2% when moving from a 4:1 to 5:1 matching solution, and is over 100% when $k > 16$ (Supporting Information Table S2).

Taken together, these simulation results highlight an important caveat: even well-matched groups, as judged by having SMD less than 0.10, can result in unacceptably large bias in estimated treatment effects. Therefore, in the scenario of a large population with relatively few treated subjects, the investigator should consider (1) imposing much stricter imbalance criteria (e.g. standardized differences <0.05) and (2) finding the optimal *k:*1 solution that is not too large (because of increased bias), but also not too small (because attrition may leave some treated individuals with no matched controls). This second point can be informed statistically. As an example, assume that the number of matched controls lost for each treated subject follows a binomial distribution with common probability of 50% (i.e. 50% attrition). If we choose to match four controls to each treated individual, then the expected proportion of treated individuals left with no remaining controls will be $(0.50)^4 = 0.0625$. In other words, approximately 94% of the remaining treated individuals will be expected to have at least one matched control.

Conclusion

In this paper, we have described a systematic approach to identifying the optimal number of controls in matching studies; one that balances the trade-off between bias and matched sample size. This involves generating tables and graphs of summary statistics and numeric balance measures for as many *k:*1 matching solutions as the data allow, and then determining the matching solution with the largest number of matched sets that will maintains overall covariate balance. Throughout this process, investigators should consider not only global measures of imbalance but measures for individual predictors. Our simulations, echoing the findings of others, show that matching more than four controls per treated subject can lead to increasingly biased treatment effects. However in longitudinal studies, even if an investigator is satisfied with a 1:1 matching ratio, a somewhat larger ratio will protect against attrition of controls. Implementation of such an approach will increase the likelihood of accurately identifying effective interventions.

Acknowledgement

We wish to thank Julia Adler Milstein for reviewing the manuscript and providing helpful suggestions that greatly enhanced the paper.

Supporting Information

Additional Supporting Information may be found in the online version of this article at the publisher's web-site:

Table S1. Sample sizes and standardized differences of Monte Carlo simulated data
Table S2. Treatment effects and bias estimates for the Monte Carlo simulated data

References

1. Rubin, D. B. (1973) Matching to remove bias in observational studies. *Biometrics*, 29, 159–184.
2. Abadie, A. & Imbens, G. (2002) *Simple and bias-corrected matching estimators for average treatment effects*. Technical Working Paper T0283: NBER.
3. Austin, P. C. (2010) Statistical criteria for selecting the optimal number of untreated subjects matched to each treated subject when using many-to-one matching on the propensity score. *American Journal of Epidemiology*, 172, 1092–1097.
4. Rassen, J. A., Shelat, A. A., Myers, J., Glynn, R. J., Rothman, K. J. & Schneeweiss, S. (2012) One-to-many propensity score matching in cohort studies. *Pharmacoepidemiology and Drug Safety*, 21 (Suppl 2), 69–80.
5. Smith, H. (1997) Matching with multiple controls to estimate treatment effects in observational studies. *Sociological Methodology*, 27, 325–353.
6. Linden, A. (2011) Designing a prospective study when randomization is not feasible. *Evaluation & the Health Professions*, 34, 164–180.
7. Stuart, E. A. (2010) Matching methods for causal inference: a review and a look forward. *Statistical Science*, 25 (1), 1–21.
8. Caliendo, M. & Kopeinig, S. (2008) Some practical guidance for the implementation of propensity score matching. *Journal of Economic Surveys*, 22, 31–72.
9. Austin, P. C. (2009) Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. *Statistics in Medicine*, 28, 3083–3107.
10. Chambers, J. M., Cleveland, W. S., Kleiner, B. & Tukey, P. A. (1983) Graphical Methods for Data Analysis. Belmont, CA: Wadsworth.
11. Cox, N. J. (2005) Speaking Stata: density probability plots. *Stata Journal*, 5, 259–273.
12. Flury, B. K. & Reidwyl, H. (1986) Standard distance in univariate and multivariate analysis. *The American Statistician*, 40, 249–251.
13. Normand, S. L. T., Landrum, M. B., Guadagnoli, E., Ayanian, J. Z., Ryan, T. J., Cleary, P. D. & McNeil, B. J. (2001) Validating recommendations for coronary angiography following an acute myocardial infarction in the elderly: a matched analysis using propensity scores. *Journal of Clinical Epidemiology*, 54, 387–398.
14. Rubin, D. B. (2001) Using propensity scores to help design observational studies: application to the tobacco litigation. *Health Services & Outcomes Research Methodology*, 2, 169–188.
15. Cohen, J. (1988) Statistical Power Analysis for the Behavioral Sciences, 2nd edn. Hillsdale, NJ: Lawrence Erlbaum.
16. Ming, K. & Rosenbaum, P. R. (2000) Substantial gains in bias reduction from matching with a variable number of controls. *Biometrics*, 56, 118–124.
17. Austin, P. C. (2008) Assessing balance in measured baseline covariates when using many-to-one matching on the propensity score. *Pharmacoepidemiology and Drug Safety*, 17, 1218–1225.
18. Linden, A. (2011) Identifying spin in health management evaluations. *Journal of Evaluation in Clinical Practice*, 17, 1223–1230.
19. Austin, P. C. (2007) Propensity-score matching in the cardiovascular surgery literature from 2004 to 2006: a systematic review and suggestions for improvement. *Journal of Thoracic Cardiovascular Surgery*, 134, 128–1135.
20. Austin, P. C. (2008) A critical appraisal of propensity-score matching in the medical literature between 1996 and 2003. *Statistics in Medicine*, 27, 2037–2049.
21. Sturmer, T., Joshi, M., Glynn, R. J., Avorn, J., Rothman, K. J. & Schneeweiss, S. (2006) A review of the application of propensity score methods yielded increasing use, advantages in specific settings, but not substantially different estimates compared with conventional multivariable methods. *Journal of Clinical Epidemiology*, 59, 437–447.
22. Rosenbaum, P. R. & Rubin, D. B. (1983) The central role of propensity score in observational studies for causal effects. *Biometrika*, 70, 41–55.
23. Lunt, M. (2007) *FGMATCH: stata module to perform 1:k greedy matching*. Available at: http://personalpages.manchester.ac.uk/staff/mark.lunt/fgmatch.ado (last accessed 17 July 2013).
24. Parsons, L. S. (2004) Performing a 1:N case-control match on the propensity score. *Proceedings of the Twenty-Ninth Annual SAS® User Group International Conference, Montreal, Canada*: SAS Institute, Inc. Available at: http://www2.sas.com/proceedings/sugi29/165-29.pdf (last accessed 17 July 2013).
25. Lunt, M. (2007) *PBALCHK: stata module to check covariate balance after matching, weighting or stratifying*. Available at: http://personalpages.manchester.ac.uk/staff/mark.lunt/pbalchk.ado (last accessed 17 July 2013).
26. Austin, P. C. (2009) Some methods of propensity-score matching had superior performance to others: results of an empirical investigation and Monte Carlo simulations. *Biometrical Journal*, 51, 171–184.
27. Morgan, S. L. & Todd, J. J. (2008) A diagnostic routine for the detection of consequential heterogeneity of causal effects. *Sociological Methodology*, 38, 231–281.
28. Iacus, S. M., King, G. & Porro, G. (2011) Multivariate matching methods that are monotonic imbalance bounding. *Journal of the American Statistical Association*, 106, 345–361.
29. Linden, A. (2006) What will it take for disease management to demonstrate a return on investment? New perspectives on an old theme. *American Journal of Managed Care*, 12, 217–222.
30. Rubin, D. B. (1973) The use of matched sampling and regression adjustment to remove bias in observational studies. *Biometrics*, 29, 185–203.