Statistics and Its Interface Volume 11 (2018) 51–60

On the use of balancing scores and matching in testing for exposure effect in case-control studies

Roni Eilenberg and Ruth Heller*

Balancing scores, especially the propensity score, are widely used to adjust for measured confounders in prospective studies. In case-control studies, the distribution of the exposure and outcome given the covariates is distorted when there is an exposure effect, due to the selection process. Therefore, it is less obvious how to estimate balancing scores. Extensive simulations revealed several interesting findings on the use of estimated balancing scores in testing for exposure effect. First, that with the aid of an estimated balancing score obtaining matched sets with a low absolute standardized difference in covariate means was far easier than without the aid of an estimated balancing score. Second, that the approach for estimating the balancing score matters, and that several potential approaches result in an inflation of the type I error probability. Third, that the strategy by which we adjust for difference in estimated balancing scores across groups can have a great effect on the power of the test for exposure effect. In particular, in our simulations, the full matching strategy on cavariates and on the estimated balancing score resulted in better power than the strategy of covariate adjustment or inverse probability weighting. We show the usefulness of full matching with the aid of our recommended approach to estimating the balancing score in a case-control study.

AMS 2000 subject classifications: Primary 62P10.
Keywords and phrases: Hypothesis testing, Overt bias, Propensity score, Retrospective studies, Stratification score.

1. INTRODUCTION

In observational studies, the distribution of covariates may differ in exposed and unexposed subjects. Confounding can be controlled by comparing exposed and unexposed subjects with the same value of confounding covariates. If there are many potential confounders, exact matching on the covariates is difficult or impossible. However, matching on the scalar propensity score (PS), i.e., the probability of exposure given the covariates, is sufficient to remove bias due to all observed covariates [15]. The PS is a balancing score, i.e., upon conditioning on the PS covariates are independent of the exposure, and moreover it is the coarsest balancing score [15].

*Corresponding author.

A widely used strategy for testing for exposure effect in an observational study is as follows [17]. First, estimate a balancing score. Second, match exposed and unexposed people that are close on the estimated balancing score, as well as on the actual covariate values (see, e.g., Chapter 8 in [18] for multivariate matching approaches). Third, check for covariate balance by comparing the absolute standardized difference in covariate means of exposed and unexposed subjects. If the balance is not satisfactory, revisit the first two steps to improve balance. Finally, test the null hypothesis of no association between exposure and outcome in matched sets. The attractiveness of this strategy is that the analysis is similar to the analysis in a blocked randomized experiment, since the outcome is only used at the final testing stage.

In a case-control study, cases are deliberately over-represented, and controls are under-represented. This design is commonly employed when the disease is rare [3]. Typically, all the cases are sampled, along with a subset of controls chosen from a population of controls (possibly using frequency matching on key covariates). Often, additional covariates are collected on the sampled individuals, and exact matching is not possible unless the additional covariates are very few. However, due to the selection process, the estimated PS may not correspond to the PS in the target population [9, 1].

The above matching strategy can be used without the aid of a balancing score (i.e., the starting point is the second step where matching is done using the covariate values) [19, 20]. However, few studies have found that using balancing scores in case-control studies for estimation of the exposure effect can be useful [5, 9, 13]. In particular, the estimated stratification score (SS), i.e., the estimated probability of the disease given the covariates, was suggested for case-control genetic association studies [5]. The SS is a retrospective balancing score for a case-control study [1], and therefore it may play in case-control studies the role that the PS plays in prospective studies. Reversing the roles of exposure and outcome in the above strategy, results in an approach in which the exposure is only used at the final testing stage.

In this work we assess different potential tests for exposure effect in case-control studies with the aid of estimated balancing scores. We focus on the common setting in which exact matching on the covariates measured only on the sampled individuals is not possible. We report the results of simulation studies, from which we can draw several important

conclusions for studies of typical size. We show the usefulness of our preferred test in a real data example from the case-control study of [11], where a questionnaire was administered to the sampled individuals in order to collect additional covariates.

2. MATERIALS AND METHODS

2.1 Methods for estimating balancing scores in retrospective studies

We use the following notation: E is the binary exposure status, D is the binary disease status, $\vec{X}$ is the vector of covariates. The PS is $e(\vec{X}) = P(E = 1 \mid \vec{X})$. The SS is $d(\vec{X}) = P(D = 1 \mid \vec{X})$. We denote A independent of B conditional on C as $A \perp B \mid C$. Let S be the selection indicator. We assume that the selection of cases and controls depends on D only, so that $S \perp (E, \vec{X}) \mid D$. (One may more generally assume that the selection also depends on the covariates used for selecting the control sample.)

We consider the following PS estimation approaches, suggested in [9]:

Subcohort (method B in [9]) A subcohort is randomly chosen from the complete cohort at the outset of the study, as opposed to case-control studies in which a subset of controls is chosen from the population of controls [14]. The PS is estimated using only the subcohort, and therefore the estimated PS is consistent if the PS model is correctly specified [9].

Weighted (method C in [9]) Estimating the PS from the entire sample, giving the controls weights that are inversely proportional to their sampling fraction. This method requires that the sampling fraction of controls be known, which is not true in many case-control studies. The estimated PS is consistent if the PS model is correctly specified.

Control only (method D in [9]) Estimating the PS from the sampled controls. When the null hypothesis that $E \perp D \mid \vec{X}$ is true, since $S \perp (E, \vec{X}) \mid D$, then $P(E = 1|\vec{X}, D = 0, S = 1) = P(E = 1|\vec{X})$, so the estimated PS is consistent if the PS model is correctly specified. When the outcome is rare, the covariate distribution of controls approaches that of the entire population (even if the null is false), so this estimated PS is nearly consistent.

Unweighted (method E in [9]) Estimating the PS from all observations in the study, without using weights. When the null hypothesis that $E \perp D \mid \vec{X}$ is true, since $S \perp (E, \vec{X}) \mid D$, then $P(E = 1|\vec{X}, S = 1) = P(E = 1|\vec{X})$, so the estimated PS is consistent if the PS model is correctly specified.

Modeled-control (method F in [9], known as the exposure score of Miettinen [12]) Estimating the PS from the cases and controls using the following two stage algorithm. First, fit a model for the probability of exposure given the covariates and the outcome. Second, for each subject, the estimated PS is the probability of exposure from the estimated model coefficients, treating all subjects as controls. When the null hypothesis that $E \perp D \mid \vec{X}$ is true, since $S \perp (E, \vec{X}) \mid D$, the coefficient of the outcome in the model for the probability of exposure is equal zero if the model is correctly specified, so the estimated PS is consistent. When the outcome is rare, the estimated PS is nearly consistent since it estimates the probability of exposure among the noncases (even if the null is false).

We consider the following SS estimation approaches:

Unweighted The SS is estimated using all subjects in the case-control study, by fitting a logistic regression model. When the null hypothesis that $E \perp D \mid \vec{X}$ is true, since $S \perp (E, \vec{X}) \mid D$, then the estimated SS is consistent if the data were generated from a logistic model [8].

DRS The disease risk score (DRS) is $P(D = 1 \mid \vec{X}, E = 0)$ [13]. The exposure status and all potential confounding variables are included as explanatory variables in a model to predict the outcome. Then, the estimated SS is obtained by setting the exposure status to zero in the fitted model. When the null hypothesis that $E \perp D \mid \vec{X}$ is true, since $S \perp (E, \vec{X}) \mid D$, then $P(D = 1|\vec{X}, E = 0, S = 1) = P(D = 1|\vec{X}, S = 1)$, so the estimated SS is consistent if the data were generated from a logistic model [8].

2.2 Methods for testing using the estimated balancing score

We considered the following methods.

Stratification For prospective studies, a common recommendation is to stratify into quintiles of the estimated PS, following the observation in [4] that under certain conditions it removes 90% of the bias. For the study sizes considered in our simulations, to remove the necessary bias for the probability of type I error to be controlled at the nominal level, stratification into quintiles was not enough, but stratification into 25 quantiles of the estimated balancing score by the unweighted method was. The association between the exposure E and the binary outcome D on strata formed using the estimated balancing score was tested using the Mantel-Haenszel test [10].

Covariate Adjustment The most common way the estimated PS is used in the analysis of observational studies in medical research is via covariate adjustment [2]. This method adjusts for imbalances by using both the estimated PS and exposure indicator (E) as explanatory variables in the logistic regression with the disease status D as outcome. The association between the exposure E and the binary outcome D was tested using the estimated regression coefficient.

IPTW This method adjusts for imbalances by using inverse probability of treatment (i.e., exposure) weighting (IPTW) in a logistic model for the outcome, including no predictors for the outcome other than exposure. The weight for a subject is $E/\hat{e}(\vec{X}) + (1 - E)/(1 - \hat{e}(\vec{X}))$, where $\hat{e}(\vec{X})$ is the estimated PS. The association between the exposure E and the binary outcome D was tested using the estimated regression coefficient.

Full Matching The most common matching procedure is pair matching, and an optimal pair matching procedure

pairs subjects with $E = 1$ ($D = 1$) with subjects with $E = 0$ ($D = 0$), so that the total distance within matched pairs is minimized. We used the more general optimal full matching algorithm, available from CRAN in the R package *fullmatch* [7]. This algorithm minimizes the total distance, among all possible partitions of the subjects to strata, so that in each stratum there is exactly one subject with $E = 1$ ($D = 1$) and at least one subject with $E = 0$ ($D = 0$), or exactly one subject with $E = 0$ ($D = 0$) and at least one subject with $E = 1$ ($D = 1$). Full matching was recommended in [6] over other matching techniques for observational studies, by showing that it was more successful in removing bias due to observed covariates. Unlike in the stratification method, the number of matched sets is determined by the data. Unlike all previous methods, this method also uses the covariates themselves, not just the balancing score. The pairwise distance matrix required as input for the full-matching on E (D) algorithm was computed as suggested in [18]:

1. Replaces each of the covariates, one at a time, by its ranks, with average ranks for ties. Let $\vec{\tilde{X}}_j$ denote the vector of ranks of the covariates for the jth subject, $j = 1, \ldots, N$.
2. Premultiply and postmultiply the empirical covariance matrix of $\vec{\tilde{X}}_j, j = 1...N$, by a diagonal matrix whose diagonal elements are the ratios of the standard deviation of untied ranks, to the standard deviations of the tied ranks of the covariates. Let $\hat{\Sigma}$ be the resulting matrix.
3. The (i, j)th entry in the distance matrix, where the rows are the subjects with E (D) value of one and the columns are the subjects with E (D) value of zero, is $(\vec{\tilde{X}}_i - \vec{\tilde{X}}_j)^T \hat{\sum}^{-1} (\vec{\tilde{X}}_i - \vec{\tilde{X}}_j)$.
4. Let $b(\vec{X})$ denote the balancing score: $\hat{e}(\vec{X})$ if matching on E, and $\hat{d}(\vec{X})$ if matching on D. When the difference in balancing score $b(\vec{X})$ between subjects i and j was greater than w, the (i, j)th entry in the distance matrix was modified to be the original distance plus a penalty (specifically, $1000 \times (|b(\vec{x}_i) - b(\vec{x}_j)| - w)$ in our simulations). In order to assess the impact of the balancing scores in full matching, we compared $w = 0$ with $w > 0$ (specifically, w was 5% of the standard deviation of $b(\vec{X})$). We call the procedure full-matching with estimated balancing score as a caliper when $w > 0$, and full-matching with no caliper when $w = 0$.

On the resulting matched sets, the association between the exposure E and the binary outcome D on strata formed using the estimated balancing score was tested using the Mantel-Haenszel test [10].

Stratification and Full Matching were considered using either the PS or the SS as the balancing score, but Covariate Adjustment and IPTW were only considered using the PS as the balancing score. Although it is possible to use Covariate Adjustment and IPTW with the SS as the balancing score, we did not consider this in our paper since these adjustment methods are not commonly used with the SS as the balancing score, and in limited simulations (not shown) Stratification and Full Matching using the SS score had far better power.

2.3 Simulation design

We conducted a series of simulations of studies of typical size to evaluate the performance, i.e., probability of type I error and power, of the different tests. We had the following main goals: to assess which balancing score estimation approach result in a probability of type I error below the nominal level, and has good power properties; to compare the four methods for testing using the estimated balancing score (full matching, stratification, covariate adjustment, and IPTW); with full matching, to compare the tests that use the estimated balancing score as a caliper to the tests with no caliper.

To simulate the cohort population we used a design similar to that of [9]. The covariates for each subject, $\vec{X}_i = (X_{1i}, \ldots, X_{10i})$, were generated independently from the standard normal distribution. Only the first p covariates influenced the exposure or the outcome. The PS and probability of outcome data generation models for subject i were:

$$\log\left(\frac{e(\vec{X}_i)}{1 - e(\vec{X}_i)}\right) = \alpha_e + \sum_{j=1}^{p} \beta_{ej} X_{ji}, \tag{1}$$

$$\log\left(\frac{P(D = 1 \mid E_i, \vec{X}_i)}{1 - P(D = 1 \mid E_i, \vec{X}_i)}\right) = \alpha_r + \sum_{j=1}^{p} \beta_{rj} X_{ji} + \log(OR) \times E_i$$

where E_i is the exposure status of subject i, drawn as an independent Bernoulli random variable with probability of exposure $e(\vec{X}_i)$. The odds ratio for the association between exposure and outcome, OR, was set to the value of one (the null case) or 1.5. In the simulation setting with a more common (less rare) disease, similar to the simulation setting of [9], the cohort size was 2,000, $\alpha_e = 0$, $\alpha_r = -log(9)$, $\beta_e = (0.2, 0.2, 0.2, 0.2, 0.2, 0, 0, 0, 0, 0)$, $\beta_r = (a, a, a, a, a, 0, 0, 0, 0, 0)$ where $a \in \{-0.6, -0.3, 0.3, 0.6\}$. In the simulation setting with a rare disease the cohort size was 20,000, $\alpha_r = -log(99)$, $\alpha_e \in \{0, -log(4)\}$, $\beta_e = (a, a, a, a, a, a, 0, 0, 0, 0)$ and $\beta_r = (-a, -a, -a, a, a, a, 0, 0, 0, 0)$ for $a \in \{-0.6, -0.3, 0.3, 0.6\}$.

After simulating the cohort population, all cases and a random sample of controls from the non-cases comprised the case-control sample. The sampling fraction of controls was fixed at 20% in the more common disease setting, and at 2% in the rare disease setting.

We conducted additional simulations to assess the sensitivity of the different methods on test performance when the data generation model had a nonlinear relationship with one of the covariates, or an unobserved confounder. Specifically, to the right hand side of both equations in (1), we added the term mX_{1i}^2 or the unobserved confounder mU_i, where the U_is were generated independently from the standard normal distribution, and $m \in \{\pm 0.1, \pm 0.2, \pm 0.3\}$.

Our results are based on 4,000 datasets generated for each parameter setting.

2.4 The study on exposure to POP and TGCTs

In [11] the effect of POP exposure on the risk of TGCTs was examined using the data from the US Servicemen's Testicular Tumor Environmental and Endochrine Determinants (STEED) Study. To be eligible for the study, cases had to be 45 years or younger at the time of diagnosis and to have donated at least one serum sample between January 1, 1987 and December 31, 2002, to the Department of Defense Serum Repository (DoDSR). Men with a serum sample in the DoDSR who had not developed TGCT were eligible to participate as controls. Each case was matched to all available potential control subjects on birth year (within 1 year), ethnicity (white, black, other), and date of available serum sample (within 30 days), using the computerized Defense Medical Surveillance System (DMSS) database. From the list of possible controls, four men were chosen at random as the control set, and they were contacted sequentially so that one control will be enrolled for each case (for some cases, the sequential scheme of contacting potential controls resulted in more than one control per case).

For illustration, we show here our analysis of the 927 controls, and 318 cases of seminoma (one of the two histological subgroup considered in [11]). We test for exposure effect of the seven POPs considered in [11] on the risk of seminoma. As in [11], we consider the subject exposed if it is in the fourth quartile of the POP distribution, and unexposed if it is in the first quartile of the POP distribution. As in [11], we adjusted for age at blood donation, ethnicity, date of serum draw, age at reference date, personal history of cryptorchism, family history of TGCT, height, and BMI. The latter four covariates were collected on the sampled individuals via questionnaires. Our caliper was the estimated SS, $\hat{d}(\vec{X})$, and it was re-estimated using logistic regression for each POP.

3. RESULTS

3.1 Simulation results

3.1.1 Results on the probability of type I error

Table 1 shows the significance level for the different tests considered. Applying the full matching algorithm is not enough for removing overt bias if a caliper is not used, since the probability of type I error is far above the nominal 0.05 level in some settings (rows 1 and 8). However, if the true PS or SS caliper is used, the nominal 0.05 level is controlled (rows 2 and 9). This suggests that applying the full matching algorithm may be enough also if the PS or SS are estimated from the data.

Interestingly, when applying the full matching algorithm, the only estimation method which resulted in a valid test, i.e., a test with probability of type I error at most the nominal level, was the unweighted method (rows 6 and 10). Estimating the caliper from the entire dataset with weighting or using the modeled controls or DRS approach resulted in a small inflation of the type I error probability (rows 4, 7, and 11). A larger inflation was observed when estimating the caliper by a subsample, i.e. by the subcohort or controls only (rows 3 and 5). While this inflation was smaller than without using any caliper, it was still high, reaching above 0.1 in some of the settings.

When stratifying on the true PS or SS, there was a small inflation of the type I error probability when adjusting by 5 strata in the less rare disease scenario (rows 12 and 19), but no inflation when adjusting by 25 strata. We further considered the validity of the tests when stratifying into 25 strata on the estimated PS or SS. As with the full-matching adjustment method, the only estimation method which resulted in a valid test, was the unweighted method (rows 17 and 21). The other methods resulted in inflations similar in magnitude to those observed with the full-matching adjustment method.

With covariate adjustment and IPTW, there was no inflation of the type I error probability when adjusting for the true PS (rows 23 and 29), or the PS estimated by the unweighted method (rows 27 and 33). All other estimation methods resulted in inflations, which were typically higher than the inflations for the same estimation methods observed when adjusting by full-matching or stratification.

An informal diagnostic check that the overt bias has been controlled is through examination of whether balance was achieved in all covariates [18]. Table 2 counts the number of datasets that were unbalanced, where a dataset is counted as unbalanced if the absolute standardized difference in covariate means is greater than 0.25 for at least one covariate. We see that the worst balance by far is when matching without a caliper. Using estimated balancing scores we achieve better covariate balance than using the true balancing scores. This is because adjusting for the estimated score removes both systematic and chance imbalances, while adjusting for the true balancing score removes only systematic imbalances, as noted by [9]. The balance based on the estimated PS was typically better than the balance based on the estimated SS, because the overlap between the estimated PS of exposed and non-exposed was larger than the overlap between the estimated SS of the cases and controls. The threshold of 0.25 may be too liberal for part of the estimation methods, which appear balanced although they are not. Specifically, although it looks like the balance achieved by the modeled-control method is only slightly worse than by the unweighted method, the difference in balance between the methods for large $|a|$ is manifest when using a lower threshold than 0.25 (not shown).

3.1.2 Results on test power

In Table 3 we show the subset of methods that controlled the 0.05 significance level. Among the oracle procedures,

Table 1. In the null case, the probability of type I error for adjustment methods: Full Matching (Rows 1–11), Stratification (Rows 12–22), Covariate Adjustment (Rows 23–28), and IPTW (Rows 29–34) (Standard Error ≤ 0.0034). The simulation parameters were: in columns 4–7 cohort size 2,000, sample fraction of controls 20%, $\alpha_e = 0$, $\alpha_r = -log(9)$, $\beta_e = (0.2, 0.2, 0.2, 0.2, 0.2, 0, 0, 0, 0, 0)$, $\beta_r = (a, a, a, a, a, 0, 0, 0, 0, 0)$; in columns 8–11 cohort size 20,000, sample fraction of controls 2%, $\alpha_r = -log(99)$, $\alpha_e = 0$, $\beta_e = (a, a, a, a, a, a, 0, 0, 0, 0)$ and $\beta_r = (-a, -a, -a, a, a, a, 0, 0, 0, 0)$

			Disease less rare, $a =$				Disease more rare, $a =$			
Row	Adjustment method	Caliper	-0.6	-0.3	0.3	0.6	-0.6	-0.3	0.3	0.6
1	Full matching on E	None	0.000	0.002	0.365	0.782	0.053	0.046	0.052	0.060
2	Full matching on E	True PS	0.037	0.040	0.047	0.042	0.044	0.043	0.047	0.049
3	Full matching on E	$\widehat{PS}$ Subcohort	0.021	0.028	0.074	0.094	0.093	0.059	0.056	0.098
4	Full matching on E	$\widehat{PS}$ Weighted	0.028	0.028	0.062	0.075	0.089	0.055	0.054	0.096
5	Full matching on E	$\widehat{PS}$ Controls only	0.020	0.026	0.081	0.143	0.102	0.058	0.054	0.107
6	Full matching on E	$\widehat{PS}$ Unweighted	0.036	0.038	0.045	0.042	0.028	0.036	0.038	0.029
7	Full matching on E	$\widehat{PS}$ Modeled control	0.045	0.043	0.053	0.057	0.075	0.056	0.056	0.083
8	Full matching on D	None	0.000	0.004	0.257	0.806	0.047	0.045	0.049	0.057
9	Full matching on D	True SS	0.035	0.039	0.048	0.042	0.040	0.049	0.048	0.051
10	Full matching on D	$\widehat{SS}$ Unweighted	0.038	0.036	0.048	0.047	0.028	0.034	0.039	0.032
11	Full matching on D	$\widehat{SS}$ DRS	0.039	0.040	0.050	0.046	0.069	0.050	0.056	0.075
12	Stratifying into 5	true PS	0.026	0.032	0.060	0.064	0.041	0.040	0.046	0.044
13	Stratifying into 25	true PS	0.038	0.041	0.045	0.043	0.041	0.041	0.046	0.044
14	Stratifying into 25	$\widehat{PS}$ Subcohort	0.020	0.030	0.072	0.102	0.092	0.058	0.053	0.101
15	Stratifying into 25	$\widehat{PS}$ Weighted	0.023	0.030	0.064	0.080	0.086	0.054	0.051	0.096
16	Stratifying into 25	$\widehat{PS}$ Controls only	0.018	0.024	0.086	0.156	0.101	0.057	0.056	0.110
17	Stratifying into 25	$\widehat{PS}$ Unweighted	0.034	0.037	0.047	0.040	0.018	0.038	0.034	0.020
18	Stratifying into 25	$\widehat{PS}$ Modeled control	0.044	0.041	0.052	0.058	0.068	0.054	0.054	0.075
19	Stratifying into 5	true SS	0.026	0.033	0.060	0.064	0.035	0.042	0.048	0.046
20	Stratifying into 25	true SS	0.038	0.042	0.045	0.043	0.035	0.044	0.046	0.047
21	Stratifying into 25	$\widehat{SS}$ Unweighted	0.039	0.037	0.048	0.046	0.019	0.035	0.034	0.026
22	Stratifying into 25	$\widehat{SS}$ DRS	0.041	0.040	0.053	0.049	0.064	0.052	0.054	0.072
23	Covariate adjustment	True PS	0.044	0.050	0.056	0.045	0.050	0.050	0.053	0.055
24	Covariate adjustment	$\widehat{PS}$ Subcohort	0.079	0.052	0.060	0.082	0.145	0.072	0.065	0.150
25	Covariate adjustment	$\widehat{PS}$ Weighted	0.065	0.051	0.058	0.071	0.141	0.073	0.064	0.140
26	Covariate adjustment	$\widehat{PS}$ Controls only	0.126	0.060	0.073	0.127	0.163	0.076	0.066	0.163
27	Covariate adjustment	$\widehat{PS}$ Unweighted	0.038	0.044	0.049	0.044	0.016	0.038	0.033	0.018
28	Covariate adjustment	$\widehat{PS}$ Modeled control	0.059	0.052	0.058	0.062	0.108	0.064	0.066	0.110
29	IPTW	True PS	0.051	0.044	0.050	0.053	0.052	0.047	0.051	0.054
30	IPTW	$\widehat{PS}$ Subcohort	0.086	0.060	0.061	0.080	0.147	0.072	0.080	0.145
31	IPTW	$\widehat{PS}$ Weighted	0.066	0.057	0.055	0.071	0.138	0.072	0.078	0.139
32	IPTW	$\widehat{PS}$ Controls only	0.156	0.078	0.078	0.151	0.159	0.074	0.082	0.159
33	IPTW	$\widehat{PS}$ Unweighted	0.022	0.038	0.038	0.021	0.023	0.035	0.039	0.028
34	IPTW	$\widehat{PS}$ Modeled control	0.092	0.062	0.063	0.096	0.101	0.062	0.068	0.100
	Average number of cases		313	232	231	314	511	260	258	510
	Average number of controls		337	354	354	337	389	396	394	389

which assume the true PS or true SS is known, the full matching on E and stratification into 25 strata using the true PS had highest power (rows 1 and 5). The full matching on D and stratification into 25 strata using the true SS had less power in the more rare disease scenario (rows 3 and 7). Covariate adjustment and IPTW also had lower power (rows 9 and 11).

Using the unweighted estimation method, the covariate adjustment and IPTW methods (rows 10 and 12) had lower power than full-matching (rows 2 and 4) and stratification (rows 6 and 8).

Full-matching and stratification, using the unweighted estimation method, had similar power, and this power was lower than the power of the tests that use the true (yet unknown in practice) balancing score in some of the settings in the rare disease simulation. The reduction in power was by at most 20% by using the estimated balancing score instead of the true balancing score. The loss of power is due

Table 2. The number of unbalanced datasets out of 4000, when the odds ratio is 1, for the two adjustment methods: Full Matching (rows 1–11) and Stratification (rows 12–22). The simulation parameters were: in columns 4–7 cohort size 2,000, sample fraction of controls 20%, $\alpha_e = 0$, $\alpha_r = -log(9)$, $\beta_e = (0.2, 0.2, 0.2, 0.2, 0.2, 0, 0, 0, 0, 0)$, $\beta_r = (a, a, a, a, a, 0, 0, 0, 0, 0)$*; in columns 8–11 cohort size 20,000, sample fraction of controls 2%,* $\alpha_r = -log(99)$, $\alpha_e = 0$, $\beta_e = (a, a, a, a, a, a, 0, 0, 0, 0)$ *and* $\beta_r = (-a, -a, -a, a, a, a, 0, 0, 0, 0)$

			Disease less rare, $a =$				Disease more rare, $a =$			
Row	Adjustment method	Caliper	-0.6	-0.3	0.3	0.6	-0.6	-0.3	0.3	0.6
1	Full matching on E	None	1859	2192	1984	1498	4000	3805	3804	4000
2	Full matching on E	True PS	38	71	69	39	636	119	123	622
3	Full matching on E	$\widehat{PS}$ Subcohort	4	3	4	4	805	49	58	839
4	Full matching on E	$\widehat{PS}$ Weighted	0	3	1	1	785	43	55	784
5	Full matching on E	$\widehat{PS}$ Controls only	17	11	17	18	881	53	56	844
6	Full matching on E	$\widehat{PS}$ Unweighted	0	0	0	0	263	5	8	259
7	Full matching on E	$\widehat{PS}$ Modeled control	0	1	0	0	285	6	10	283
8	Full matching on D	None	4000	3108	3041	4000	4000	3382	3385	4000
9	Full matching on D	True SS	1379	153	187	1339	1688	87	77	1620
10	Full matching on D	$\widehat{SS}$ Unweighted	412	2	1	413	632	3	4	628
11	Full matching on D	$\widehat{SS}$ DRS	420	2	1	415	661	2	6	650
12	Stratifying into 5	true PS	68	122	144	85	572	176	184	514
13	Stratifying into 25	true PS	90	188	195	125	831	249	242	849
14	Stratifying into 25	$\widehat{PS}$ Subcohort	7	5	14	20	1231	90	104	1238
15	Stratifying into 25	$\widehat{PS}$ Weighted	2	5	6	6	1164	81	95	1153
16	Stratifying into 25	$\widehat{PS}$ Controls only	44	28	56	80	1300	99	109	1300
17	Stratifying into 25	$\widehat{PS}$ Unweighted	0	0	1	0	184	5	9	167
18	Stratifying into 25	$\widehat{PS}$ Modeled control	0	1	0	0	285	6	10	283
19	Stratifying into 5	true SS	915	265	257	941	1271	186	194	1227
20	Stratifying into 25	true SS	1337	308	312	1310	1328	242	232	1333
21	Stratifying into 25	$\widehat{SS}$ Unweighted	265	3	0	247	384	3	2	398
22	Stratifying into 25	$\widehat{SS}$ DRS	253	3	1	246	418	1	1	423

Table 3. The power when the odds ratio is 1.5, for: Full Matching (rows 1–4), Stratification (rows 5–8), Covariate Adjustment (rows 9–10), IPTW (rows 11–12). The simulation parameters were: in columns 4–7 cohort size 2,000, sample fraction of controls 20%, $\alpha_e = 0$, $\alpha_r = -log(9)$, $\beta_e = (0.2, 0.2, 0.2, 0.2, 0.2, 0, 0, 0, 0, 0)$, $\beta_r = (a, a, a, a, a, 0, 0, 0, 0, 0)$*; in columns 8–11 cohort size 20,000, sample fraction of controls 2%,* $\alpha_r = -log(99)$, $\alpha_e = 0$, $\beta_e = (a, a, a, a, a, a, 0, 0, 0, 0)$ *and* $\beta_r = (-a, -a, -a, a, a, a, 0, 0, 0, 0)$. *(Standard Error* ≤ 0.008*)*

			Disease less rare				Disease more rare			
Row	Adjustment method	Caliper	$a = -.6$	$a = -.3$	$a = .3$	$a = .6$	$a = -.6$	$a = -.3$	$a = .3$	$a = .6$
1	Full matching on E	True PS	0.650	0.687	0.700	0.658	0.725	0.762	0.754	0.732
2	Full matching on E	$\widehat{PS}$ Unweighted	0.644	0.686	0.693	0.652	0.568	0.692	0.686	0.569
3	Full matching on D	True SS	0.641	0.683	0.692	0.664	0.623	0.716	0.708	0.637
4	Full matching on D	$\widehat{SS}$ Unweighted	0.640	0.685	0.688	0.646	0.558	0.688	0.676	0.555
5	Stratifying into 25	true PS	0.660	0.700	0.704	0.664	0.732	0.770	0.765	0.740
6	Stratifying into 25	$\widehat{PS}$ Unweighted	0.643	0.685	0.700	0.647	0.526	0.696	0.678	0.536
7	Stratifying into 25	true SS	0.660	0.698	0.704	0.659	0.530	0.676	0.668	0.546
8	Stratifying into 25	$\widehat{SS}$ Unweighted	0.666	0.698	0.711	0.661	0.536	0.690	0.680	0.528
9	Covariate adjustment	True PS	0.583	0.637	0.630	0.596	0.660	0.713	0.698	0.664
10	Covariate adjustment	$\widehat{PS}$ Unweighted	0.564	0.622	0.626	0.570	0.423	0.611	0.601	0.431
11	IPTW	True PS	0.544	0.622	0.635	0.517	0.544	0.622	0.635	0.517
12	IPTW	$\widehat{PS}$ Unweighted	0.470	0.583	0.584	0.407	0.470	0.583	0.584	0.407
	Average number of cases		356	272	282	368	618	321	320	620
	Average number of controls		329	345	344	327	387	394	394	387

Table 4. The number of unbalanced datasets out of 4000, when the odds ratio is 1.5, for the two adjustment methods: Full Matching (rows 1–4), and Stratification (rows 5–8). The simulation parameters were: in columns 4–7 cohort size 2,000, sample fraction of controls 20%, $\alpha_e = 0$, $\alpha_r = -log(9)$, $\beta_e = (0.2, 0.2, 0.2, 0.2, 0.2, 0, 0, 0, 0, 0)$, $\beta_r = (a, a, a, a, a, 0, 0, 0, 0, 0)$*; in columns 8–11 cohort size 20,000, sample fraction of controls 2%,* $\alpha_r = -log(99)$, $\alpha_e = 0$, $\beta_e = (a, a, a, a, a, a, 0, 0, 0, 0)$ *and* $\beta_r = (-a, -a, -a, a, a, a, 0, 0, 0, 0)$

			Disease less rare				Disease more rare			
Row	Adjustment method	Caliper	$a = -.6$	$a = -.3$	$a = .3$	$a = .6$	$a = -.6$	$a = -.3$	$a = .3$	$a = .6$
1	Full matching on E	True PS	76	134	189	113	857	242	241	825
2	Full matching on E	$\widehat{PS}$ Unweighted	0	0	0	0	225	5	2	240
3	Full matching on D	True SS	1193	104	150	1413	1531	53	52	1563
4	Full matching on D	$\widehat{SS}$ Unweighted	285	3	0	425	538	3	3	545
5	Stratifying into 25	true PS	76	134	189	113	857	242	241	825
6	Stratifying into 25	$\widehat{PS}$ Unweighted	0	0	0	0	141	3	3	184
7	Stratifying into 25	true SS	1164	226	274	1308	1206	188	198	1222
8	Stratifying into 25	$\widehat{SS}$ Unweighted	172	2	6	285	356	5	4	371

Table 5. The probability of type I error when the data generation model is misspecified. Specifically, to the right hand side of both models in equation (1), we added the term mX_{1i}^2 *(columns 4–10) or the unobserved confounder* mU_i *(columns 11–16), where the* U_i*s were generated independently from the standard normal distribution, and* $m \in \{\pm 0.1, \pm 0.2, \pm 0.3\}$*. The caliper was estimated using the unweighted method. The other simulation parameters were: cohort size 20,000, sample fraction of controls 2%,* $\alpha_r = -log(99)$, $\alpha_e = 0$, $\beta_e = (0.3, 0.3, 0.3, 0.3, 0.3, 0.3, 0, 0, 0, 0)$ *and* $\beta_r = (-0.3, -0.3, -0.3, 0.3, 0.3, 0.3, 0, 0, 0, 0)$

Adjustment method, and Caliper	Nonlinear covariate term mX_{1i}^2						Unobserved confounder term mU_i					
	-0.3	-0.2	-0.1	0.1	0.2	0.3	-0.3	-0.2	-0.1	0.1	0.2	0.3
Full matching on E,$\widehat{PS}$	0.092	0.057	0.040	0.042	0.058	0.097	0.090	0.066	0.051	0.052	0.108	0.340
Full matching on D,$\widehat{SS}$	0.095	0.060	0.042	0.042	0.059	0.089	0.104	0.067	0.051	0.048	0.094	0.281
Stratifying into 25,$\widehat{PS}$	0.087	0.051	0.036	0.036	0.048	0.087	0.085	0.060	0.046	0.045	0.104	0.323
Stratifying into 25,$\widehat{SS}$	0.086	0.051	0.038	0.036	0.052	0.084	0.088	0.060	0.046	0.044	0.102	0.310

to the fact that when the $OR > 1$, the estimated balancing score using the unweighted method is biased because of the case-control sampling. Therefore, the evidence of association between D and E within the strata or matched sets defined by this biased estimate is weaker than the evidence if defined by the true balancing score.

Table 4 show the number of unbalanced datasets for the valid tests. We see that the number of unbalanced datasets is significantly smaller when using the estimated balancing score (by the unweighted method), instead of the true balancing score.

3.1.3 Sensitivity of type I error probability to model misspecification and an unobserved confounder

Table 5 shows the results for the unweighted estimation method of the balancing score, testing using stratification or full-matching with the estimated balancing score as caliper. The probability of the type I error was at most the nominal level for $|m| = 0.1$, but it was inflated for $|m| > 0.1$. As expected, the further m is from zero the greater the inflation, since the bias of the estimated balancing score increases with $|m|$. Out of 4000 datasets, at most 7, 9, 9, 8, 18, and 61 datasets were unbalanced when m was $-0.3, -0.2, -0.1.0.1, 0.2$, and 0.3, respectively, clearly indicating that the standard diagnostic check of covariate balance we employed cannot account for misspecification by addition of a quadratic term, nor can it account for bias due to an unobserved covariate (as expected). More elaborate diagnostic tests (which do not only check mean differences), would be able to identify the misspecification due to the added term mX_{1i}^2, and a sensitivity analysis is necessary in order to assess the sensitivity of a significant result to unobserved confounding.

3.2 The study on exposure to persistent organochlorine pesticides (POP) and testicular germ cell tumors (TGCTs)

Table 6 shows the P values using three testing methods. After Bonferroni correction at level 0.05: full matching without caliper resulted in three significant findings (an exposure effect of Cis-nonachlor, p, p'-DDE, and Trans-nonachlor); full matching with caliper (our recommended method) resulted in two findings (an exposure effect of Cis-nonachlor and Trans-nonachlor); stratification resulted in one finding (an exposure effect of Cis-nonachlor). Table 7 shows that the balance is better for exposures Cis-nonachlor, p, p'-DDE,

Table 6. The P value of the Mantel-Haenszel test of asscociation of Seminoma with each exposure (exposed if in fourth quartile, unexposed if in first quartile, of the POP Level) for the two adjustment methods: Stratification on the estimated SS by the unweighted method (column 1), and Full-Matching (Column 2 with estimated SS as caliper, column 3 without caliper)

Exposure	Stratifying into 25	Full matching	
(Q4 vs Q1)	$\hat{SS}$ Unweighted	caliper $\hat{SS}$ Unweighted	no caliper
Cis-nonachlor	0.0043	0.0028	5e-04
p,p'-DDE	0.0426	0.0205	0.0015
p,p'-DDT	0.2136	0.1682	0.1007
HCH	0.6212	0.513	0.0658
Mirex	0.5059	0.6334	0.3545
Oxychlordane	0.0463	0.034	0.0097
Trans-nonachlor	0.0102	0.0031	0.0052

Table 7. The standardized difference in covariate mean between the cases and controls after adjustment by matching/stratification, for three exposures and the three adjustment methods: Stratification using the estimated SS by the unweighted method, Full-Matching using the estimated SS as caliper, Full-Matching without caliper

Exposure	Adjustment method	age at blood draw	age of serum	ref age	height (inches)	bmi	is white	is black	crypt. history	family history missing	family history yes	family history unknown
Cis-nonachlor	before	0.40	-0.00	0.40	0.20	0.10	0.10	-0.10	0.10	-0.00	0.20	0.10
	strat.$\widehat{SS}$	0.02	-0.07	0.03	0.03	0.02	-0.01	0.03	0.05	0.02	-0.01	-0.01
	fullmatch $\widehat{SS}$	0.06	-0.07	0.04	0.06	0.00	-0.08	0.14	0.01	-0.01	0.00	0.02
	fullmatch no caliper	0.18	-0.02	0.20	0.12	0.01	-0.07	0.06	-0.01	-0.16	0.09	0.02
p,p'-DDE	before	0.40	-0.20	0.40	0.30	0.10	-0.00	-0.20	0.00	0.10	0.10	0.10
	Strat.$\widehat{SS}$	-0.02	-0.03	-0.01	0.06	0.07	0.01	0.02	0.01	-0.04	-0.05	0.04
	fullmatch $\widehat{SS}$	-0.04	0.00	-0.01	0.01	0.05	-0.05	0.03	0.02	0.02	-0.05	0.05
	fullmatch no caliper	0.18	-0.09	0.21	0.16	0.04	-0.11	0.01	-0.07	-0.05	0.02	0.03
trans-nonachlor	before	0.30	-0.10	0.40	0.20	0.10	0.10	-0.10	0.20	-0.00	0.20	0.20
	strat.$\widehat{SS}$	0.06	-0.06	0.05	0.02	-0.01	0.01	-0.04	0.01	0.03	-0.04	0.04
	fullmatch $\widehat{SS}$	0.09	-0.10	0.09	-0.04	-0.04	0.00	0.04	0.01	-0.04	-0.09	0.10
	fullmatch no caliper	0.12	-0.00	0.13	0.09	0.01	-0.06	0.05	0.12	-0.06	0.04	0.02

and Trans-nonachlor when the estimated SS is used for adjustment, than with full-matching without caliper. In particular, for exposure p,p'-DDE, the full-matching without caliper results in imbalances above 0.15 for the age at reference date, date of serum draw, and height, yet the full-matching with estimated SS as caliper result in imbalances of at most 0.05.

Using our recommended method of full-matching with caliper, we can conclude that increased exposure to Cis- and Trans- nonachlor may increase the risk of seminoma. A sensitivity analysis should be conducted in order to assess how robust these conclusions are to (hidden) bias, see [16] for a useful approach.

4. DISCUSSION

Our work was motivated by the fact that while testing with the aid of estimated balancing scores is a highly useful strategy in prospective studies, there is lack of consensus on whether this strategy should be used in case-control studies, and if so how. Consistent estimation of the PS and the SS is possible if there is no exposure effect. For testing for exposure effect in case-control studies, we reached the following conclusions about the role of the PS and of the SS, based on our empirical investigations.

Our first conclusion is that an estimated balancing score is useful for stratifying, or matching, the observations prior to testing. Applying the full matching algorithm without the use of a balancing score, the significance level was very high in some of the simulation settings. When we allowed the matching algorithm to omit a certain fraction of the data, the imbalances were smaller, and the significance level was closer to the nominal level (results not shown). However, these improvements were minor compared to the excellent balance and control of the probability of type I error at the nominal level we were able to achieve using the full matching algorithm with the balancing score caliper estimated using the unweighted method. With the aid of the estimated balancing score caliper, the full matching algorithm becomes an approach aimed at achieving balance. Alternatively, it is possible to consider algorithms that di-

rectly aim at achieving balance without the aid of a balancing score, as suggested in [20] for sparse nominal covariates.

Our second conclusion is that the estimation method of the balancing score matters, and the preferred method is the one that takes all the data into account when estimating the balancing score, i.e., the unweighted method. This method is consistent when there is no effect. Other methods, which are also consistent when there is no effect, nevertheless failed to maintain the nominal significance level. The reason, pointed out in [9], is that using the unweighted method there is no artifactual effect modification. Using all other estimation methods, when the exposure does not affect the outcome, an artifactual association within the strata or matched sets of the estimated balancing score may be present, and lead to an inflated probability of type I error. To understand this artifactual association, [9] considered the simplified setting of the PS being fixed (i.e., not varying with covariates, as in a randomized trial). Then by estimating the PS using controls only, in the lowest (highest) stratum of estimated PS, the fraction of exposed within controls will be lower (higher) than the true PS, yet within cases it will be on average equal to the true PS.

Our third conclusion is that full matching with the estimated balancing score as caliper has some advantages over stratification on the estimated balancing score, covariate adjustment, or IPTW. The performance of full matching and stratification (into 25 strata) was very similar, and their power was superior to that of covariate adjustment and IPTW. Stratification has a limitation that the number of strata should be decided prior to testing, and it is not clear what the best number is with relation to sample size: if the number of strata is too small, the probability of the type I error may be inflated, and if the number of strata is too large power may be reduced. Full matching has two main advantages over stratification: the number of matched sets is determined by the data, and subjects are matched based on the distance between covariate values in addition to proximity in the estimated balancing score.

In summary, for typical study sizes considered in this paper, we have demonstrated that assuming the only bias is overt bias, a test with probability of type I error at most the nominal level can be obtained if we use the full matching algorithm with the balancing score estimated using the unweighted method as caliper. We also showed that the other estimation methods resulted in tests with inflated type I error probabilities. Even though the unweighted method does not consistently estimate the balancing score when the null hypothesis is false, the power loss from using the estimated balancing score instead of the true balancing score was small. In the real data example, we achieved a far better balance in the covariates of matched sets with the aid of the estimated balancing score than without using the balancing score in the full matching algorithm.

In our simulation settings, when the null hypothesis of no effect of exposure on outcome is true, the task of estimating the propensity score was similar in difficulty to the task of estimating the stratification score. In practice, it may be easier to estimate one of the balancing scores. For example, in genetic association studies, estimating the stratification score is easier than estimating the propensity score [1], and matching on D is more natural than matching on E.

ACKNOWLEDGEMENTS

The authors thank Katherine McGlynn and Barry Graubard for providing the data example. The authors also thank Malka Gorfine, Barry Graubard, Ruth Pfeiffer, Paul Rosenbaum, Dylan Small, and David Steinberg for useful discussions on earlier versions of this manuscript.

Received 5 February 2016

REFERENCES

[1] Allen, A. and Satten, G. Control for confounding in case-control studies using the stratification score, a retrospective balancing score. *American Journal of Epidemiology.* 2011; 173(7):752–760.

[2] Austin, P., Grootendorst, P., Normand S., and Anderson, G. Conditioning on the propensity score can result in biased estimation of common measures of treatment effect: A Monte Carlo study. *Statistics in Medicine.* 2007; 26:754–768. MR2339172

[3] Breslow, N. Statistics in epidemiology: The case-control study. *Journal of the American Statistical Association.* 1996; 91:14–28. MR1394064

[4] Cochran, W. The effectiveness of adjustment by subclassification in removing bias in observational studies. *Biometrics.* 1968; 24:295–313. MR0228136

[5] Epstein, M., Allen, A., and Satten, G. A simple and improved correction for population stratification in case-control studies. *American Journal of Human Genetics.* 2007; 80(5):921–930.

[6] Hansen, B. Full matching in an observational study of coaching for the SAT. *Journal of the American Statistical Association.* 2004 99(467):609–618. MR2086387

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

[8] Hosmer, D. and Lemeshow, S. *Applied Logistic Regression.* Wiley Series in Probability and Statistics, 2000. MR3287463

[9] Mansson, R., Joffe, M., Sun, W., and Hennessy, S. On the estimation and use of propensity scores in case-control and case-cohort studies. *American Journal of Epidemiology.* 2007; 166(3):332–339.

[10] Mantel, N. and Haenszel, W. Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute.* 1959; 22(4):719–748.

[11] McGlynn, K., Quraishi, S., Graubard, B., Weber, J., Rubertone, M., and Erickson, R. Persistent organochlorine pesticides and risk of testicular germ cell tumors. *J Natl Cancer Inst.* 2008; 100(9):663–671.

[12] Miettinen, O. Stratification by a multivariate confounder score. *American Journal of Epidemiology.* 1976; 104(6):609–620.

[13] Pfeiffer, R. and Riedl, R. On the use and misuse of scalar scores of confounders in design and analysis of observational studies. *Statistics in Medicine.* 2015; 34:2618–2635. MR3368406

[14] Prentice, R. A case-cohort design for epidemiological cohort studies and disease prevention trials. *Biometrika.* 1986; 73:1–11.

[15] Rosenbaum, P. and Rubin, D. The central role of the propensity score in observational studies for causal effects. *Biometrika.* 1983; 70(1):41–55. MR0742974

[16] Rosenbaum, P. Sensitivity analysis for matched case-control studies. *Biometrics.* 1991; 47:87–100. MR1108691

[17] Rosenbaum, P. *Observational Studies*. 2nd ed. Springer Series in Statistics, 2002. MR1899138
[18] Rosenbaum, P. *Design of Observational Studies*. Springer Series in Statistics, 2010. MR2561612
[19] Small, D. S., Cheng, J., Halloran, M. E., and Rosenbaum, P. R. Case definition and design sensitivity. *Journal of the American Statistical Association*. 2013; 108:1457–1468. MR3174721
[20] Zubizarreta, J. R., Reinke, C. E., Kelz, R. R., Silber, J. H., and Rosenbaum, P. R. Matching for several sparse nominal variables in a case-control study of readmission following surgery. *Journal of the American Statistical Association*. 2011; 65(4):229–238. MR2867507

Roni Eilenberg
Department of Statistics and Operations Research
Tel-Aviv University
Tel-Aviv
Israel
E-mail address: eilenberg@mail.tau.ac.il

Ruth Heller
Department of Statistics and Operations Research
Tel-Aviv University
Tel-Aviv
Israel

National Cancer Institute
Rockville, MD 20852
USA
E-mail address: ruheller@post.tau.ac.il