The International Journal of Biostatistics

Volume 8, *Issue* 1 2012 *Article* 25

Evaluating treatment effectiveness in patient subgroups: a comparison of propensity score methods with an automated matching approach

Rosalba Radice, *CLondon School of Hygiene & Tropical Medicine*
Roland Ramsahai, *Centre for Statistical Methodology, LSHTM*
Richard Grieve, *Centre for Statistical Methodology, LSHTM*
Noemi Kreif, *Centre for Statistical Methodology, LSHTM*
Zia Sadique, *Centre for Statistical Methodology, LSHTM*
Jasjeet S. Sekhon, *University of California, Berkeley*

Recommended Citation:
Radice, Rosalba; Ramsahai, Roland; Grieve, Richard; Kreif, Noemi; Sadique, Zia; and Sekhon, Jasjeet S. (2012) "Evaluating treatment effectiveness in patient subgroups: a comparison of propensity score methods with an automated matching approach," *The International Journal of Biostatistics*: Vol. 8: Iss. 1, Article 25.
DOI: 10.1515/1557-4679.1382

Evaluating treatment effectiveness in patient subgroups: a comparison of propensity score methods with an automated matching approach

Rosalba Radice, Roland Ramsahai, Richard Grieve, Noemi Kreif, Zia Sadique, and Jasjeet S. Sekhon

Abstract

Propensity score (Pscore) matching and inverse probability of treatment weighting (IPTW) can remove bias due to observed confounders, if the Pscore is correctly specified. Genetic Matching (GenMatch) matches on the Pscore and individual covariates using an automated search algorithm to balance covariates. This paper compares common ways of implementing Pscore matching and IPTW, with Genmatch for balancing time-constant baseline covariates}. The methods are considered when estimates of treatment effectiveness are required for patient subgroups, and the treatment allocation process differs by subgroup. We apply these methods in a prospective cohort study that estimates the effectiveness of Drotrecogin alfa activated, for subgroups of patients with severe sepsis. In a simulation study we compare the methods when the Pscore is correctly specified, and then misspecified by ignoring the subgroup-specific treatment allocation. The simulations also consider poor overlap in baseline covariates, and different sample sizes. In the case study, GenMatch reports better covariate balance than IPTW or Pscore matching. In the simulations with correctly specified Pscores, good overlap and reasonable sample sizes, all methods report minimal bias. When the Pscore is misspecified, GenMatch reports the least imbalance and bias. With small sample sizes, IPTW is the most efficient approach, but all methods report relatively high bias of treatment effects. This study shows that overall GenMatch achieves the best covariate balance for each subgroup, and is more robust to Pscore misspecification than common alternative Pscore approaches.

KEYWORDS: confounding, observational studies, matching, propensity score methods, subgroup analysis

Author Notes: We thank James Carpenter and Rhian Daniel (both LSHTM) for their suggestions. We also thank David Harrison and Kathy Rowan (ICNARC) for access to the data for the motivating example.

1 Introduction

Observational studies are widely used to estimate treatment effectiveness; here the major concern is confounding (Moodie and Stephens, 2010). Regression is often used to adjust for potential confounders, but if the distribution of baseline covariates does not overlap between the treatment groups, estimates may be highly sensitive to model specification (Rubin, 1997). To reduce reliance on parametric assumptions, propensity score (Pscore) methods, including stratification, matching, regression adjustment and inverse probability of treatment weighting (IPTW), are widely used to estimate treatment effects (Austin, 2008a; Shah et al., 2005; Stürmer et al., 2006; Austin 2008b; Austin and Laupacis, 2011). Of these approaches, matching and IPTW can perform relatively well (Austin, 2009a), and IPTW has been extended to allow for time-varying exposures and confounders (Robins et al., 2000). The Pscore specification must be considered by examining covariate balance after matching or weighting, and if the resultant balance is poor, the Pscore re-estimated. However, studies rarely follow this careful process, they often fail to assess covariate balance and may report biased estimates of treatment effectiveness based on misspecified Pscores (Austin, 2008a).

Policy-makers require unbiased estimates of the average treatment effect (ATE), not just for an overall population but also for particular subgroups (Hasford et al., 2010). For studies that aim to report treatment effects for each subgroup, correct Pscore specification is particularly challenging (Lefebvre and Gustafson, 2010). A major concern is that each subgroup may have a different treatment assignment mechanism. Reliable inference then requires that the Pscore balances baseline characteristics across treatment groups within each subgroup. A Pscore approach has to then recognize the differential treatment assignment mechanism, for example by estimating separate Pscore models for each subgroup. If the Pscore is misspecified, because for example the same Pscore is used for each subgroup, then the treatment groups will be imbalanced. Hence IPTW or Pscore matching may provide incorrect inferences (Drake, 1993).

Instead of relying on correct Pscore specification, covariate balance can be achieved with multivariate matching methods that attempt to directly balance individual characteristics, for example within each subgroup of interest. Genetic matching (GenMatch) combines Pscore matching with multivariate matching on the individual covariates, using an automated search algorithm to optimize covariate balance (Diamond and Sekhon, 2012; Sekhon, 2011). GenMatch can reduce bias and mean squared error (MSE) compared to Pscore matching (Diamond and Sekhon, 2012; Sekhon, 2011), and has been applied across a diverse range of settings (Gilligan and Sergenti, 2008; Gordon and Huber, 2007; Grieve et al., 2008; Heinrich, 2008; Herron and Wand, 2007; Korkeamaki and Uusitalo, 2009; Lenz and

Ladd, 2009; Woo et al., 2008). Alternative automated approaches include using targeted maximum-likelihood estimation (e.g., van der Laan and Gruber, 2010; van der Laan, 2010a; van der Laan, 2010b). However, none of these papers compares GenMatch to IPTW, or reports treatment effectiveness for subgroups.

This paper aims to compare GenMatch to common ways of implementing Pscore matching and IPTW for tackling confounding, when reporting treatment effectiveness by subgroup. The methods are considered in a motivating example and a simulation study. The motivating example assesses the effectiveness of a controversial pharmaceutical intervention, Drotrecogin alfa activated (DrotAA) for severe sepsis, the most common cause of death in adult intensive care units (ICUs) (Rowan et al., 2008). The Protein C Worldwide Evaluation in Severe Sepsis (PROWESS) trial reported that DrotAA reduced overall 28-day mortality versus placebo (Bernard et al., 2001), but posthoc subgroup analysis suggested benefit solely for high-risk patients. These findings generated the hypothesis that the effectiveness of DrotAA may differ according to baseline severity. We compare alternative Pscore approaches in re-analyzing a previous observational study estimating the effectiveness of DrotAA (Rowan et al., 2008). Each approach uses the previously published Pscore, which after matching, gave reasonable levels of covariate balance across the treatment groups (Rowan et al. 2008) according to conventional standards (Austin 2008a). Unlike the previous study we recognize the differential treatment allocation by subgroup. We then conduct a simulation study that extends the motivating example, and examines the relative bias and precision following each Pscore approach, when the subgroup-specific treatment allocation is recognized, and then ignored. We also consider settings with baseline covariates that have poor overlap between the treatment groups, and according to different sample sizes.

2 Methods

2.1 Statistical methods

The methods considered all assume that confounding can be removed by balancing observed baseline covariates, and require choices to be made in advance, about which variables are potential confounders. Variables should be chosen for inclusion in the Pscore or matching algorithm, so as to balance potential confounders. The choice should not be based on statistical tests for baseline differences (Rubin, 2008), but can draw on theory, published literature, expert opinion or causal diagrams (Pearl, 1995).

For each statistical method we estimate ATEs. This estimand can be obtained for matching methods by matching both a control observation to each patient in the treatment group, and a treated observation to each observation in the control group (Abadie et al., 2001), and this is the standard estimand for IPTW. The ATEs were reported for the same populations of interest, represented by the distribution of characteristics across both treatment groups in the unmatched data (Kurth et al., 2006). For the matching approaches treated and control individuals were matched to their nearest neighbor in the comparison group, one-to-one, with replacement (Abadie and Imbens, 2009).

We report the treatment effects with a common measure, the marginal odds ratio (OR). Marginal effects have high policy relevance as they apply to the population or subpopulation of interest, whereas conditional effects refer to the individual. Except under certain restrictive settings (Greenland et al., 1999) marginal and conditional ORs differ, i.e. the OR is non-collapsible (Austin, 2007). For IPTW, we weight observed outcomes for both treatment and control groups (Robins et al., 2000). For both matching approaches, we calculate ORs across all the matched pairs (Abadie et al., 2001). Given concerns about the interpretability of ORs, in the case study we also report treatment effects as relative risks, using Poisson regression.

An important challenge for the statistical methods is that the treatment assignment mechanism may differ by subgroup. For example, the relative influence of factors explaining treatment assignment may differ for high risk versus low risk patients. Balancing baseline characteristics for overall samples of treated and control observations can leave potential confounders imbalanced at the subgroup-level. In this context the methods aim to achieve covariate balance at the subgroup level.

We used weighted standardized differences to assess covariate balance which is a recommended measure for comparing balance between IPTW and matching methods (Austin, 2009b). Here the matching uses frequency weights, and IPTW the inverse of the Pscore, to weight the means and variances of the covariates (Austin, 2009b). Some researchers suggest a standardized difference of 10% denotes meaningful imbalance (Austin, 2009b; Normand et al., 2001), others that balance should be maximized without limit (Sekhon, 2011; Imai et al., 2008).

2.1.1 Propensity score matching

Assuming no unobserved confounding and that the distributions of baseline covariates overlap between the treatment groups (Cole and Hernán, 2008), matching on a correctly specified Pscore can balance observed covariates and reduce bias (Rosenbaum and Rubin, 1983). To check the Pscore specification, covariate balance should

be examined, and if balance is poor, the Pscore re-estimated (Rosenbaum and Rubin, 1984; Stuart, 2010). If the Pscore model is misspecified, balance may not be achieved and the Pscore matching estimator is biased and inconsistent (Sekhon, 2011). In particular, where ATEs are required for subgroups, and the treatment assignment mechanism differs by subgroup, matching on the Pscore estimated across the whole sample may not balance covariates in each subgroup of interest. Instead, for each subgroup separate Pscores can be estimated, and used to create subgroup-specific matched datasets.

2.1.2 Genetic matching - matching on the Pscore and individual covariates

Rosenbaum and Rubin (1985) recommended combining matching on the Pscore with matching on individual covariates, using the Mahalanobis distance (MD). This approach improves balance if the covariates follow ellipsoidal distributions, such as the normal (Rubin, 1992). However, in practice, this approach can lead to worse covariate balance, for example in the presence of binary variables, and in finite samples (Sekhon, 2011).

GenMatch can combine matching on the Pscore and covariates, but rather than selecting matched pairs according to their closeness, this approach optimizes covariate balance between the matched treatment and control samples. GenMatch selects matched pairs using a generalized MD metric, which includes an additional vector of weights for each covariate included in the matching. The weights define different distance metrics, which differ in the relative importance given to matching on each covariate. An automated search algorithm selects those weights (Sekhon and Mebane, 1998; Mebane and Sekhon, 2011), and hence the corresponding distance metric, that gives the best covariate balance in the matched samples. The choice of balance statistic has to be made *a priori* from recommended traditional measures, such as standardized mean differences, or more general measures such as Kolmogorov-Smirnov (KS) tests and empirical quantile plots (Austin, 2009b). Balance can be optimized separately for the subgroups of interest, and treatment effects can be reported using separate matched datasets for each subgroup.

A general practical concern is that if a covariate chosen for the Pscore or GenMatch algorithm is not associated with outcome, then conditioning on this covariate will increase variance without reducing bias in the estimated treatment effect (Austin et al., 2007; Brookhart et al., 2006; Schisterman et al., 2007). If the GenMatch algorithm is required to balance unnecessary covariates (Schisterman et al., 2007), this increases the dimensionality of the matching problem, which with a small sample size can increase the bias inherent in multivariate matching estimators (Abadie and Imbens, 2006). More details on GenMatch are given in Appendix A.

2.1.3 Inverse probability of treatment weighting

IPTW estimates treatment effectiveness by using the Pscore to weight the treatment and control samples (Hernán, 2000; Hirano et al., 2003; Lunceford and Davidian 2004). The weight, w_i is the inverse of the estimated probability of the observed treatment, that is $w_i = \frac{T_i}{\hat{\pi}_i} + \frac{1-T_i}{1-\hat{\pi}_i}$, where $\hat{\pi}_i$ is the estimated Pscore for the i-th individual and T_i is the treatment indicator. Individuals with a high predicted probability of the observed treatment receive a relatively low weight. When there is good overlap and the Pscore model is correctly specified, the IPTW estimator can provide unbiased and relatively efficient estimates of the ATE (Hirano et al., 2003). However, even with good overlap, if the Pscore is misspecified, baseline covariates can be imbalanced, which can lead to bias and inefficiency (Pearl, 1995). With poor overlap, the weights can be extreme which can lead to increased bias and variance (Pearl, 1995; Rosenbaum and Rubin 1983). Here, a recommended strategy is to truncate the weights (Cole and Hernán, 2008). When there are different treatment assignment mechanisms for each subgroup, the weights can be taken from separate Pscores estimated for each subgroup.

2.2 Description of motivating example

A prospective cohort study previously matched patients who received DrotAA to controls, and reported that DrotAA was effective for high-risk (three to five organ failures at baseline), but not for low-risk patients (two organ failures) (Rowan et al., 2008). However, this study did not consider whether potential confounders were balanced for each subgroup. Our reanalysis included in the Pscore the same baseline covariates, reported in Table 1, as the original study $(\text{sample size } n = 2,726)$. To address confounding we extended the previous study and recognized that treatment allocation may differ by subgroup.

The Pscore methods initially used the original Pscore model, common across both patient subgroups (overall Pscore), and the GenMatch algorithm was required to improve covariate balance across the whole sample (overall GenMatch algorithm). The analyses were repeated but with separate Pscore models (subgroup-specific Pscore) and GenMatch algorithms (subgroup-specific GenMatch algorithm) for each subgroup defined *a priori* according to whether patients had two, or three to five organ failures (see Appendix B). Here, the matching methods created separate matched datasets for each subgroup.

For each method covariate balance was reported for those baseline factors which *a priori* were judged potential confounders. This list of variables for assessing balance differed from the set of variables in the Pscore and matching algorithm

(see Table 1). Expert opinion was used to designate which of these variables were *high*, or *low priority* variables to balance (Sadique et al., 2011). The most important confounders were anticipated to be age, the proportion ventilated at ICU admission (% Ventilated), the acute physiology score (IMscore), and the baseline probability of death, calculated as a function of 20 underlying physiological variables (IMprob). To balance these confounders, it was judged necessary to include in the Pscore, and matching algorithm each of the baseline covariates listed above, but not all these variables were designated potential confounders. Some of the covariates in the Pscore and matching algorithm were not regarded as important to balance themselves, but were included to help balance the major confounders (Sadique et al., 2011). Of the variables designated as being of some importance, the GenMatch algorithm was required to maximise balance on the *high* and then the *low priority* variables. For further details on this approach to prioritising the covariates to balance see Ramsahai et al. (2011).

To address imbalances beyond differences in means, matching methods can also use non-parametric KS tests (Diamond and Sekhon, 2012; Stuart 2010). As a sensitivity analysis, the GenMatch algorithm was modified to optimize balance assessed by KS and t-tests.

Marginal ORs were estimated by logistic regression applied to each matched dataset, and for IPTW the logistic regression incorporated weights calculated from each Pscore (Sekhon, 2011; Stuart, 2010). As well as reporting ORs, we also reported relative risks. A sensitivity analysis was performed for IPTW by truncating the weights. There were no missing data.

2.3 Motivating example results

Here we present balance for those baseline covariates that were anticipated to be the major confounders: age, IMprob, IMscore and % ventilated. The standardized differences before matching were large for both subgroups (Table 2), and there was reasonable overlap (Appendix C, Figure 3).

Following Pscore matching and IPTW, some large standardized differences remained for either subgroup, whether using the overall Pscore or the subgroup-specific Pscore (Table 3). GenMatch reported better balance than the other methods when required to balance across the overall sample for one subgroup (3 to 5 organ failures), but for the other subgroup (2 organ failures) none of the methods was dominant in terms of covariate balance. Balance improved further when the algorithm balanced at the subgroup level. When subgroup-specific GenMatch algorithms were applied, GenMatch achieved better balance on the *high priority* variables than Pscore matching and IPTW and similar balance on the *low priority*

Table 1: Variables included in the Pscore and in the balance matrix.

Variable	Pscore	Balance matrix: *high priority*	Balance matrix: *low priority*
Age	✓	✓	✗
IMscore	✓	✓	✗
% Ventilated	✓	✓	✗
Pre-existing conditions:	✓	✗	✓
Number of organ system failing	✓	✗	✓
Sex	✓	✗	✗
Types of organ system failure (cardiovascular, respiratory, renal, hematological, and metabolic acidosis)	✓	✗	✗
Number of critical care beds	✓	✗	✗
Source of admission to critical care	✓	✗	✗
Diagnostic category	✓	✗	✗
IMprob	✗	✓	✗
Organ system failing in first 24 hours:			
Card/Resp[a]	✗	✗	✓
Card/Resp/Acid	✗	✗	✓
Card/Resp/Renal/Acid	✗	✗	✓

[a] Card/Resp is the abbreviation for cardiovascular/respiratory organ system failure, Card/Resp/Acid for cardiovascular/respiratory/ metabolic acidosis and Card/Resp/Renal/Acid for cardiovascular/respiratory/renal/metabolic acidosis.

Table 2: Baseline characteristics for DrotAA and control patients.

Covariate	DrotAA	Controls	Standardized difference (%)[a]
2 organ failures subgroup	(n=198)	(n=630)	
Age	57.58	63.04	26.49
IMprob	0.42	0.39	10.76
IMscore	22.83	20.44	29.53
% Ventilated	88.38	70.16	40.02
3 to 5 organ failures subgroup	(n=878)	(n=1,020)	
Age	58.96	65.16	32.32
IMprob	0.64	0.58	20.12
IMscore	32.08	27.96	40.83
% Ventilated	93.39	78.53	38.90

[a] Continuous variables are reported as means, dichotomous variables as proportions. Note absolute standardized differences are reported as percentages.

variables (see Tables 8 and 9, Appendix C). In the sensitivity analysis when KS statistics were included in the GenMatch optimization, GenMatch again reported improved balance for *high priority* variables and similar balance for *low priority* variables compared to the other approaches (see Appendix C, Tables 10, 11 and 12). When the IPTW weights were truncated above the first and 99*th* percentiles, covariate balance worsened (see Appendix C, Table 13).

Table 3: Covariate balance of *high priority* variables when using i) an overall Pscore or GenMatch algorithm and ii) a subgroup-specific Pscore or GenMatch algorithm. Results reported are weighted standardized differences (%)[a].

		Overall Pscore or GenMatch		Subgroup-specific Pscore or GenMatch	
Covariate	Method	2 organ failures group	3 to 5 organ failures group	2 organ failures group	3 to 5 organ failures group
Age					
	Pscore matching	5.00	1.17	0.93	1.25
	GenMatch	0.76	0.38	0.37	0.01
	IPTW	6.63	2.31	9.06	7.24
IMprob					
	Pscore matching	2.77	1.54	15.97	6.82
	GenMatch	1.21	0.31	0.37	0.01
	IPTW	0.56	0.82	5.58	9.86
IMscore					
	Pscore matching	8.83	6.96	6.82	3.95
	GenMatch	13.11	3.56	0.35	0.01
	IPTW	13.48	3.85	3.88	12.41
% Ventilated					
	Pscore matching	3.91	1.87	8.33	2.48
	GenMatch	5.01	2.66	0.23	0.00
	IPTW	12.29	0.57	5.99	13.19

[a] Note absolute standardized differences are reported as percentages.

Table 4 reports marginal ORs for the effect of DrotAA versus control on hospital mortality. The corresponding relative risks are reported in Appendix C (Table 14). The effectiveness of DrotAA differed by subgroup; the CIs for the treatment by subgroup interactions excluded zero (see Table 15, Appendix C).

For the two organ failures subgroup, the point estimates all exceeded 1, but for IPTW the CIs were wide especially after weighting with the overall Pscore (Table 4). GenMatch reported similar ORs whether the algorithm was required

to match across the overall sample or for each subgroup. When GenMatch was required to optimize balance according to KS and t-tests, the estimated treatment effects were similar to the base case.

Table 4: Effectiveness of DrotAA versus controls for each subgroup with: (i) an overall Pscore or GenMatch algorithm and (ii) a subgroup-specific Pscore or GenMatch algorithm. Results are reported as marginal ORs (95% CIs[a]) of hospital mortality.

	Pscore matching	GenMatch	IPTW
i) Overall Pscore and GenMatch			
2 organ failures subgroup	$1.77\ (1.49, 2.03)$	$1.60\ (1.26, 1.91)$	$1.82\ (0.77, 2.80)$
3 to 5 organ failures subgroup	$0.70\ (0.63, 0.76)$	$0.63\ (0.54, 0.70)$	$0.70\ (0.51, 0.88)$
ii) Subgroup-specific Pscore and GenMatch			
2 organ failures subgroup	$1.90\ (1.55, 2.22)$	$1.56\ (1.24, 1.86)$	$1.55\ (0.86, 2.16)$
3 to 5 organ failures subgroup	$0.64\ (0.57, 0.71)$	$0.60\ (0.52, 0.68)$	$0.78\ (0.43, 1.11)$

[a] Confidence intervals (CIs) were calculated by bootstrapping. Inference after matching should follow recent recommendations and be regarded as conditional on the estimated Pscore and the matched data (Stuart, 2010).

2.4 Simulation description

We conducted Monte Carlo simulations to examine the relative performance of each method for estimating treatment effects by subgroup. The three scenarios considered were grounded in the motivating example, and prior concerns about each method. The first scenario misspecified the Pscore and the GenMatch algorithm by ignoring the subgroup specific treatment allocation, as in the motivating example. The second scenario considered poor overlap (see Figure 4, Appendix C), and the omission of a non-linear term from the Pscore models and GenMatch algorithms. The last scenario included a covariate not associated with outcome in the Pscore models and GenMatch algorithms (Austin et al., 2007), and considered smaller sample sizes ($n = 1,000$; $n = 100$) (Brookhart et al., 2006; Schisterman et al., 2007).

Data generating process

We used a similar data generating process (DGP) to previous studies (Austin, 2009a; Austin 2007). For each subject, two continuous confounders X_1 and X_2, were generated from a bivariate normal distribution. A third confounder, X_3, defining the patient subgroup, was a binary variable generated from a Bernoulli distribution (see Appendix B). Treatment status, T and a binary outcome variable, Y were randomly generated from Bernoulli distributions with parameters π and p, determined by a different logistic model for each pair of scenarios (see below).

Scenario 1: subgroup-specific Pscore and GenMatch algorithm (1a) versus overall Pscore and GenMatch algorithm (1b)

For each subject, the logit of the Pscore, π, was determined by: $\text{logit}(\pi) = \ln(0.2) + 0.1X_1 + 0.2X_2 + 0.3X_3 + 0.2X_1X_3 - 0.2X_2X_3$, and the logit of the parameter for the outcome model, p, by: $\text{logit}(p) = -25 + \ln(2)T + 10X_1 + 0.5X_2 + 0.2X_3 - \ln(1.5)X_3T$. The interaction terms allowed the confounders to have a differential effect on treatment assignment according to subgroup (Pscore model), and allowed treatment effects to differ by subgroup (outcome model).

In the first scenario (1a), correctly specified subgroup-specific Pscores were used for matching and IPTW weights. Similarly, GenMatch was required to match and balance on X_1, X_2 and the estimated linear predictor of π, separately for each subgroup. In scenario 1b, the Pscore and the GenMatch algorithm were both misspecified; the Pscore was estimated across both subgroups; GenMatch was required to match and balance on X_1, X_2 and X_3 across the whole sample, and the overall linear predictor of π. In these scenarios there was good overlap in the distribution of the covariates and the Pscore between the treatment groups.

Scenario 2: poor overlap, correct specification of the Pscore and GenMatch algorithms (2a) and then misspecification by exclusion of a squared term (2b)

All methods correctly attempted to maximize balance at the subgroup level, but this scenario considered poor overlap and misspecification of the treatment allocation mechanism by exclusion of a nonlinear term. The logit of the Pscore was given by: $\text{logit}(\pi) = \ln(0.1) - 0.4X_1 + 0.8X_2 + 1.2X_3 - 0.2X_1^2X_3$, and the logit of the parameter for the binary outcome model was: $\text{logit}(p) = -14 + 0.1T + 3X_1 + 0.5X_2 - 0.2X_3 + X_3T$. The means and standard deviations of the confounders X_1, and X_2 were chosen to ensure poor overlap in the distribution of the Pscore and the key confounder X_1, especially for the subgroup $X_3 = 1$ (for a comparison of overlap between Scenario 1 and 2, see Appendix C, Figure 4).

In scenario 2a we assumed a correctly specified Pscore, and GenMatch was asked to maximize balance on each confounder, including the nonlinear term X_1^2. In scenario 2b, the treatment allocation models were misspecified for the $X_3 = 1$ subgroup by excluding X_1^2.

Scenario 3: recommended exclusion of a covariate not associated with outcome (3a) versus inclusion of a covariate not associated with outcome (3b)

This scenario assumed good overlap, but the Pscore included a continuous normal covariate, X_4, not associated with the outcome: $\text{logit}(\pi) = \ln(0.2) + 0.1X_1 + 0.2X_2 + 0.3X_3 + 0.2X_1X_3 - 0.2X_2X_3 + 0.6X_4 - 0.2X_4X_3$. The outcome model was as in scenario 1.

In scenario 3a, X_4 was excluded from the estimated Pscore, and from the terms GenMatch was asked to balance. In scenario 3b, X_4 was included in the estimated Pscores and GenMatch algorithms.

We report marginal ORs as in the empirical example. Here the true marginal ORs were obtained by a Monte Carlo simulation with $10,000$ samples of size $10,000$ (Austin, 2007). In scenarios 1 and 3, these ORs were 1.105 ($X_3 = 0$) and 1.052 ($X_3 = 1$), and for scenario 2 they were 1.035, and 1.496. Recall that, owing to the non-collapsibility of the OR, the marginal ORs do not coincide with the conditional ORs (Austin, 2007). Each scenario was run with $1,000$ replications, each with a sample size of $2,000$. Scenario 3 was also run with smaller sample sizes ($n = 1,000; 100$). For all scenarios we calculated the bias and root mean squared error (RMSE) of the estimated treatment effects. For sample `R` code for the simulation see Appendix B.

2.5 Simulation study results

Table 5 reports the weighted standardized differences for scenarios 1 and 2. With good overlap and correctly specified methods (scenario 1a), the standardized differences were small. When the Pscore model was misspecified by fitting an overall Pscore, and GenMatch failed to match and balance at the subgroup level (scenario 1b), both Pscore methods had high standardized differences compared to GenMatch. Under scenario 2, with weak overlap for the $X_3 = 1$ subgroup (Figure 4, Appendix C), all methods reported worse covariate balance. The deterioration in balance was least for GenMatch and most for IPTW, here even with a

correctly specified Pscore (scenario 2a) the nonlinear term, X_1^2 was highly imbalanced (standardized difference of 24). Under scenario 2b, the Pscore for the $X_3 = 1$ subgroup excluded the nonlinear term X_1^2; balance on this term deteriorated for each method, but remained worst for IPTW (standardized differences: 25, IPTW; 21, Pscore matching and 14 GenMatch).

Table 5: Covariate balance[a] in the Monte Carlo simulation for scenarios[b] 1 and 2. Results reported are weighted standardized differences (%)

		Subgroup $X_3 = 0$		Subgroup $X_3 = 1$	
Scenario	Method	X_1	X_2	X_1	X_2
1a					
	Pscore matching	2.58	1.39	0.62	3.10
	GenMatch	0.11	0.12	0.20	0.15
	IPTW	0.51	0.55	0.80	0.59
1b					
	Pscore matching	7.86	8.51	8.30	8.50
	GenMatch	1.03	1.39	1.15	1.31
	IPTW	8.09	8.14	8.54	8.12
2a					
	Pscore matching	4.05	2.60	8.00	13.2
	GenMatch	0.80	1.34	6.79	7.18
	IPTW	2.63	3.42	19.37	10.55
2b					
	Pscore matching	3.89	2.47	12.00	12.24
	GenMatch	0.84	1.13	8.25	3.52
	IPTW	2.59	3.41	16.43	8.88

[a] Weighted standardized differences are means across the $1,000$ simulations, and are reported as percentages.
[b] The sample size in these scenarios was $2,000$.

The left (right) panels of Figure 1 report bias and variation when the Pscores

and the GenMatch algorithms are correctly (incorrectly) specified by recognizing (ignoring) the subgroup-specific treatment allocation (scenario 1). With correct specification, all methods reported treatment effects centered on their true values. Under misspecification, the estimated ORs were biased and more variable for Pscore matching and IPTW; for the $X_3 = 1$ subgroup, the relative biases were 13% (Pscore matching), 14% (IPTW) and 1% (GenMatch). The corresponding RMSEs were 0.20 (Pscore methods), and 0.08 (GenMatch).

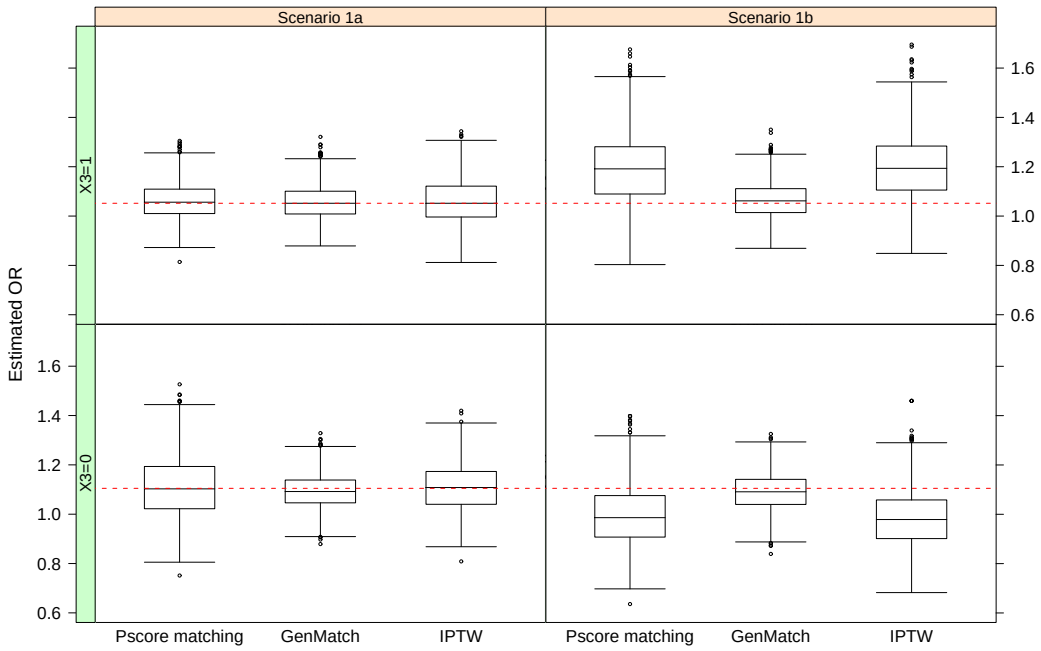

Figure 1: Boxplots showing bias and variation for the estimated ORs across $1,000$ replications for scenario 1 in the simulation study. The results in the left panel are from when the Pscore model and the GenMatch algorithm are correctly specified by recognizing the subgroup-specific treatment allocation (scenario 1a). The results in the right panel are for when the methods do not recognize the subgroup-specific treatment allocation (scenario 1b). The dashed lines are the true values.

Figure 2 reports bias and variation for scenario 2, where overlap is poor. Under correct specification, Pscore matching and GenMatch reported moderate bias (8% and 3% for $X_3 = 1$). For IPTW, where covariate balance was poor for the X_1^2 term, bias was higher (15%, for $X_3 = 1$). The corresponding RMSE for IPTW was six times that for GenMatch. With misspecification, the biases were higher for

each method (32% for IPTW, 21% for Pscore matching and 5% for GenMatch). To examine the IPTW weights scenario 2a was repeated with the same DGP but for a single dataset of $1,000,000$ (Appendix C, Figure 5). For subgroup $X_3 = 1$, the weights for the treatment group are extreme which may explain the excessive bias and variance. When the IPTW weights were progressively truncated, the standardized differences increased (Appendix C, Table 16), and the IPTW estimator became less variable, but more biased (Appendix C, Figure 6).

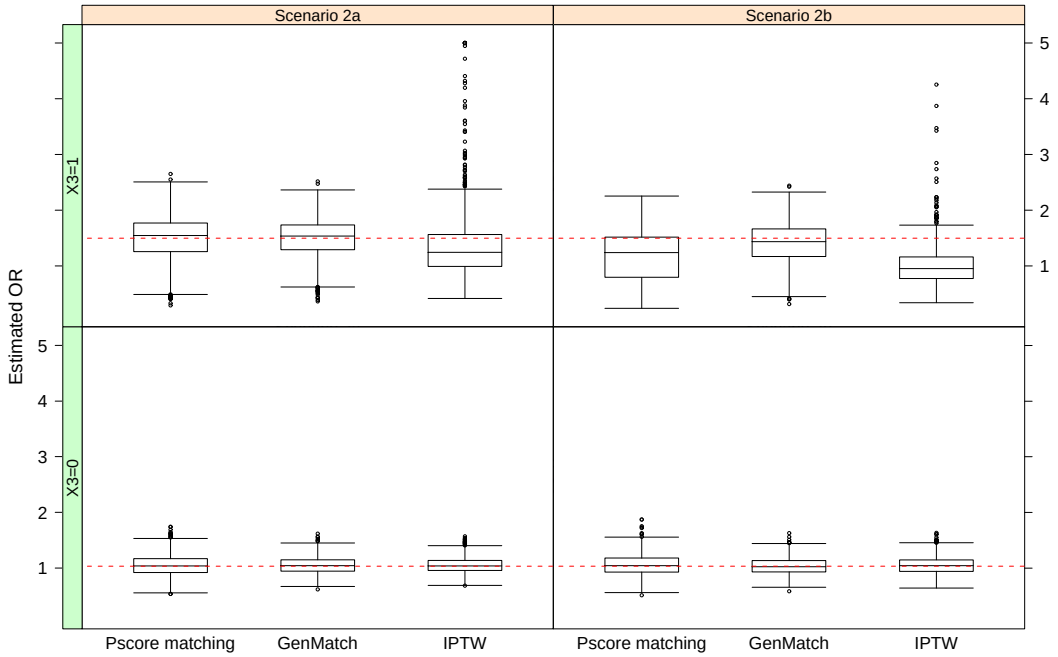

Figure 2: Boxplots showing bias and variation for the estimated ORs across $1,000$ replications for the scenario (2) with poor overlap. The left panel provides results for when the Pscore model and the GenMatch algorithm are correctly specified by including a nonlinear term (scenario 2a), the right panel for when the nonlinear term is omitted (scenario 2b). The dashed lines are the true values.

Table 6 reports the weighted standardized differences for scenario 3. With sample sizes of $2,000$ or $1,000$, the standardized differences for the true confounders (X_1 and X_2) remained small even if the methods were required to balance the covariate not associated with outcome (X_4). When the sample size was reduced to 100, balance on the confounders deteriorated especially following Pscore matching.

Table 6: Covariate balance in the Monte Carlo simulation for scenario 3 with different sample sizes[a]. Results reported are weighted standardized differences[b] (%)

		Subgroup $X_3 = 0$			Subgroup $X_3 = 1$		
Scenario	Method	X_1	X_2	X_4	X_1	X_2	X_4
3a	*n=2,000*						
	Pscore matching	2.87	1.58	58.77	0.60	2.84	39.41
	GenMatch	0.15	0.17	58.77	0.15	0.11	39.48
	IPTW	0.85	0.89	58.61	0.58	0.43	39.47
3b							
	Pscore matching	4.46	4.22	1.88	2.56	3.22	1.92
	GenMatch	0.77	0.70	0.93	0.32	0.33	0.37
	IPTW	2.34	2.35	3.20	0.99	0.84	1.13
3a	*n=1,000*						
	Pscore matching	3.97	2.47	58.74	1.23	4.25	39.31
	GenMatch	0.29	0.33	58.19	0.27	0.20	39.26
	IPTW	1.27	1.36	58.32	0.85	0.66	39.30
3b							
	Pscore matching	6.03	6.23	2.97	3.95	4.80	3.06
	GenMatch	1.28	1.18	1.51	0.53	0.52	0.61
	IPTW	3.13	3.37	4.23	1.40	1.26	1.66
3a	*n=100*						
	Pscore matching	13.90	13.53	65.60	10.10	10.91	45.47
	GenMatch	5.21	5.41	65.12	2.81	2.30	45.66
	IPTW	9.01	9.01	62.05	4.72	4.32	42.83
3b							
	Pscore matching	20.20	20.21	19.33	13.80	13.10	13.64
	GenMatch	9.31	9.72	11.85	4.10	4.31	5.24
	IPTW	13.80	13.70	17.33	7.00	6.80	8.08

[a] Across the replications the average number of treated versus controls was 69% treated, 31% controls ($n = 2,000$), 66% treated, 34% controls ($n = 1,000$), 68% treated, 32% controls ($n = 100$).

[b] Weighted standardized differences are means across the $1,000$ simulations, and are reported as percentages.

Table 7 reports bias and RMSE for scenario 3. With sample sizes of $2,000$ or $1,000$, all methods reported estimates that were relatively unbiased and statistically efficient. With a small sample size ($n = 100$), IPTW provided the least biased, most efficient estimates, but all methods performed poorly.

3 Discussion

This paper compares GenMatch with common implementations of Pscore matching and IPTW. GenMatch is an approach that combines matching on the Pscore and the individual covariates. The study considers settings where both treatment effectiveness and the treatment assignment mechanism differ by subgroup, and the major concern is balancing time-constant covariates. The case study exemplifies a general methodological challenge, that of reporting unbiased estimates when treatment effectiveness is anticipated to differ by patient subgroup (Hasford et al., 2010; Lefebvre and Gustafson, 2010). The motivating example is in critical care, where risk adjustment is relatively advanced, and the assumption of no unmeasured confounding may be judged reasonable (Rowan et al., 2008). Here it was anticipated that because receipt of DrotAA could differ by subgroup, a 'subgroup specific Pscore' would help balance covariates in each subgroup. However, achieving covariate balance with a Pscore is a challenging process (Austin, 2008), and in this case study neither of the previously recommended Pscore methods (Austin, 2009a) is able to balance covariates within the subgroups. By contrast the approach that combines matching on the Pscore and the individual covariates, does balance potential confounders in both subgroups. This case study highlights the importance of adopting an approach that achieves balance for subgroups, to enable policy makers to identify patients who would benefit most from treatment.

The simulation study finds that if, as in the motivating example, the estimated Pscore ignores a differential treatment allocation by subgroup, estimates can be biased and inefficient. GenMatch is relatively robust to this misspecification, because it aims to directly balance potential confounders using an automated search algorithm, rather than a fixed parametric model. This paper extends previous work that reports lower MSE for GenMatch compared to Pscore matching alone, or combined with MD matching (Diamond and Sekhon, 2012; Sekhon and Grieve, 2011; Kang and Schafer, 2007). This is the first study to compare these three methods, and does so in an important context for policy makers — that of subgroup analysis.

IPTW is a common method for estimating treatment effectiveness with observational data. Unlike matching, IPTW extends to handling time-varying covariates (Robins et al., 2000), and can minimize MSE if the Pscore is correctly specified

Table 7: % Bias and RMSE in the Monte Carlo simulation for scenario 3 with different sample sizes.

		% Bias		RMSE	
Scenario	Method	$X_3 = 0$	$X_3 = 1$	$X_3 = 0$	$X_3 = 1$
3a	*n=2,000*				
	Pscore matching	1.42	0.02	0.15	0.08
	GenMatch	1.25	0.23	0.09	0.07
	IPTW	0.81	0.04	0.12	0.09
3b					
	Pscore matching	1.59	0.49	0.19	0.12
	GenMatch	4.46	1.12	0.14	0.08
	IPTW	1.13	0.05	0.14	0.09
3a	*n=1,000*				
	Pscore matching	2.40	0.79	0.22	0.12
	GenMatch	2.57	0.59	0.13	0.10
	IPTW	0.99	1.34	0.13	0.17
3b					
	Pscore matching	4.27	1.38	0.28	0.18
	GenMatch	6.74	0.01	0.21	0.12
	IPTW	2.06	1.34	0.21	0.13
3a	*n=100*				
	Pscore matching	42.08	17.03	1.97	0.90
	GenMatch	31.28	12.22	1.55	0.63
	IPTW	20.76	12.05	0.94	0.60
3b					
	Pscore matching	78.10	23.18	2.83	0.96
	GenMatch	59.87	16.08	2.28	0.87
	IPTW	43.05	14.25	1.80	0.67

(Hirano et al., 2003). However, even with good overlap, if the Pscore is misspecified, IPTW can be unreliable (Kang and Schafer, 2007; Petersen et al., 2012). IPTW can report different levels of covariate balance to Pscore matching because IPTW incorporates the Pscore in the estimator (Lee et al., 2011); while Pscore matching may only use the proximity based on the Pscore to create matched pairs (Zhao, 2008). In the simulated scenario with good overlap but a misspecified Pscore, we extend the general finding that IPTW estimates can be biased and inefficient (Kang and Schafer, 2007; Ertefaie and Stephens, 2010), to the context of subgroup analysis. Faced with poor overlap and extreme weights, we apply a recommended approach and truncate the weights (Lunceford and Davidian, 2004; Scharfstein et al., 1999), but covariate balance does not improve. Alternative ways of stabilizing the weights (Cao et al., 2009), redefining the relevant population of interest (Petersen et al., 2012), or adopting doubly robust (DR) methods (Kang and Schafer, 2007; Scharfstein et al., 1999; Robins et al., 1994; Bang and Robins, 2005; Robins et al., 2007) warrant consideration. There is much debate on the relative advantages of DR methods, especially in settings with weak overlap (Petersen et al., 2012; Robins et al., 2007) and model misspecification (Lefebvre and Gustafson, 2010). Recent developments of data-adaptive DR methods show considerable promise (van der Laan, 2010a; van der Laan and Gruber, 2010; Porter et al., 2011), and further testing across a range of applications is now warranted. While missing data is beyond the scope of this paper, recent work has extended Pscore methods to this context (Mattei, 2009; Qu and Lipkovich, 2009).

Bad overlap can also hinder matching methods, leading to poor quality matches, covariate imbalance and biased estimates of treatment effects. Faced with poor overlap, either matching method can impose calipers (Stuart, 2010), but this changes the population of interest; a strength of this study is that it compares the methods in the same population. Our simulation highlights that with a small sample size, matching can lead to biased and statistically inefficient estimates relative to IPTW (Abadie and Imbens, 2006). Multivariate matching methods such as GenMatch, that focus on balancing the individual covariates, may be particularly prone to bias and imprecision when the sample size is small (Abadie and Imbens, 2006).

This study has several limitations. Each approach assumes no unmeasured confounders, an untestable assumption which in many cases is implausible. Each method requires the analyst to choose the covariates, but also the statistics for balance assessment (Brookhart et al., 2006). In the main analysis we followed recommendations and used weighted standardized mean differences (Austin, 2009b), but more general balance statistics such as non-parametric Kolmogorov-Smirnoff tests warrant consideration and as the sensitivity analysis in the case study shows, can be considered by GenMatch (see also Diamond and Sekhon, 2012).

4 Conclusion

When the estimated Pscore is misspecified, an automated approach that combines matching on the Pscore and individual covariates can report less biased estimates of treatment effectiveness for patient subgroups than common ways of implementing Pscore matching or IPTW. The combined matching approach performs less well with small sample sizes. These findings apply to settings where treatment and potential confounders are time-constant.

Appendix A

Genetic matching

Genetic matching (GenMatch) automates the process of maximizing balance on observed covariates in the matched sample by using an evolutionary search algorithm to determine the weight each individual covariate is given. As with any matching method, GenMatch requires choices to be made *a priori* about which covariates to include in the matching and assessment of balance, and which balance statistic to use. The key innovations of GenMatch are the generalised distance metric, and the use of an iterative search algorithm to maximize covariate balance. Diamond and Sekhon (2012) and Sekhon (2011) provided full details of the method and its properties in a general context, so here we summarize the key aspects.

Selection of covariates for matching algorithm

Before matching, it is necessary to choose which potential confounders to condition on. The researcher should follow general guidance and only consider those covariates anticipated to influence the outcome (Brookhart, 2006). This selection process should also consider interaction effects as well as main effects and nonlinear terms. The choice can be informed by previous empirical analyses, expert opinion, and causal diagrams (Rubin, 2008; Pearl, 1995). The GenMatch algorithm will only use those matching variables that are pre-specified. As with any other matching method the choice of variables for balance assessment should include those anticipated to be of high prognostic importance whether or not they are included in the matching. For example, a summary prognostic measure may be excluded from the matching because it is highly correlated with the underlying covariates, and better overall balance may be achieved by just matching on the covariates. GenMatch can also be tailored to prioritise achieving covariate balance on particular covariates designated as *high priority*, for further details see Ramsahai et al. (2011).

Covariate balance statistics

A recommended statistic for checking covariate balance is the weighted standardized mean difference:

$$d = \frac{\bar{x}_{\text{treatment}} - \bar{x}_{\text{control}}}{\sqrt{\frac{s^2_{\text{treatment}} + s^2_{\text{control}}}{2}}}$$

where for continuous covariates $\bar{x}$ and s^2 denote the covariate's weighted means and variances. This balance statistic allows matching methods to be compared to IPTW, by using the appropriate weights, i.e. the frequency weights in matched datasets, and the IPTW weights calculated from the Pscore. This measure can be adapted for binary variables (Austin, 2009b).

In some circumstances, the weighted standardized mean differences are an insufficient measure of balance as they are insensitive to imbalances in aspects of the covariate distribution beyond the mean (e.g., variance, maximum, skew, kurtosis). To address imbalances beyond differences in means for linear terms, matching methods can consider standardized differences for higher order terms, but also alternative balance statistics such as Kolmogorov-Smirnov (KS) tests and empirical quantile-quantile plots (Austin, 2009b). A potential advantage of GenMatch is that it can maximise balance according to whatever balance statistic the user specifies including the more general measures listed above.

Distance metric

The Mahalanobis distance (MD) between any two observations (one from treatment and the other from control) is

$$\text{MD}(\mathbf{X}_i, \mathbf{X}_j) = \left\{ (\mathbf{X}_i - \mathbf{X}_j)^{\mathsf{T}} \mathbf{S}^{-1} (\mathbf{X}_i - \mathbf{X}_j) \right\}^{1/2} \tag{1}$$

where $\mathbf{S}$ is the sample covariance matrix of $\mathbf{X}$ and $\mathbf{X}^{\mathsf{T}}$ is the transpose of the matrix $\mathbf{X}$. Using this metric, distance between individual covariates is collapsed into a single scalar.

The Pscore can be combined with MD by, for example, including the Pscore as a variable in the $\mathbf{X}$ matrix in (1).

GenMatch generalizes the MD by including an additional weight matrix $\mathbf{W}$:

$$\text{GMD}(\mathbf{X}_i, \mathbf{X}_j, \mathbf{W}) = \left\{ (\mathbf{X}_i - \mathbf{X}_j)^{\mathsf{T}} \left(\mathbf{S}^{-1/2} \right)^{\mathsf{T}} \mathbf{W} \left(\mathbf{S}^{-1/2} \right) (\mathbf{X}_i - \mathbf{X}_j) \right\}^{1/2} \tag{2}$$

where $\mathbf{W}$ is a $k \times k$ positive definite weight matrix with k being the number of matching covariates , and $\mathbf{S}^{-1/2}$ is the Cholesky decomposition of $\mathbf{S}$. GenMatch

essentially matches by minimizing the generalized version of MD given in (2). **W** is chosen to be the weight matrix that minimizes covariate imbalance according to the balance statistics the user chooses (e.g., standardized difference, KS statistics).

The GenMatch algorithm uses the distance measure, GMD (equation 2) in which (by default) all elements of **W** are zero except down the main diagonal. The main diagonal is the vector of weights chosen by the algorithm. If each of the weights for the covariates are set equal to one and the weight for the Pscore is zero, GMD is the same as MD. That is, GenMatch will converge to the MD if that proves to be the optimal distance measure. If the Pscore contains all the information required to maximize covariate balance, the algorithm will converge to the corresponding distance metric, that is, the Pscore will be given full weight, and the other elements in **W** will be given zero weight. Hence, both Pscore and MD matching can be considered as limiting cases of GenMatch. The inclusion of individual covariates in the **X** matrix, rather than relying solely on the specification of the Pscore, helps ensure covariate balance when the Pscore is misspecified. In this sense, GenMatch is robust to misspecifications in the Pscore.

The iterative search algorithm

Here we provide an overview of the optimization algorithm. Further details are available in Sekhon and Mebane (1998) and Mebane and Sekhon (2011).

The aim of the GenMatch algorithm is to find the optimal weights, **W**, that is the weights which produce the matched sample with the best balance. GenMatch uses a genetic search algorithm to search the weight matrices **W**, where each possible vector of weights corresponds to a different distance metric as defined in equation (2). The algorithm proposes batches of weights, **W** and moves towards the batch which contains the optimal weights. Each batch is a *generation* and is used iteratively to produce a subsequent generation with better candidate **W**. The size of each generation is the *population size* (e.g., 1,000) and is constant for all generations. For each generation the sample is matched according to each metric, corresponding to each **W**, to produce as many matched samples as the population size. Balance is evaluated for each matched sample and the algorithm identifies the weights corresponding to the best balance. The generation of candidate **W**s evolves towards those containing, on average, better **W** and asymptotically converges to contain the optimal **W**: the one which maximizes balance.

The **X** matrix includes all variables which are matched on and is used to define the GMD between units. The *balance matrix* consists of columns of data for each variable used to measure balance, and by default, the balance matrix is identical to the **X** matrix. Optimization can be stopped if there is no significant

improvement in the minimum loss over a specified number of generations or it can be stopped after a fixed number of generations (e.g., 200).

Previous simulation evidence

Diamond and Sekhon (2008) conducted an extensive simulation study to compare the performance of GenMatch to other matching methods (Pscore matching, MD matching, Pscore and MD matching combined). The results showed that GenMatch produced better covariate balance in each of the settings considered. Where the Pscore was correctly specified and the covariates were multivariate normal, GenMatch dominated the other multivariate matching methods in terms of bias and MSE, and reported lower MSE than Pscore matching. When the Pscore was misspecified, GenMatch reported lower bias and MSE than the other estimators.

Sekhon and Grieve (2011) compared GenMatch to Pscore matching in a challenging setting where some covariates were discrete, and others continuous but with highly skewed distributions. The simulation reported that GenMatch achieved better covariate balance, lower bias and MSE, compared with Pscore matching.

Diamond and Sekhon (2012) compared the performance of GenMatch to Pscore matching, where the Pscore was estimated by a linear logistic regression model, random forests and boosted Classification and Regression Trees. The simulations considered scenarios that differed in the degree of linearity and additivity in the true Pscore model, that is the extent to which the Pscore model included quadratic and interaction terms. GenMatch reported the smallest MSE and bias, apart from one scenario where matching on the correctly specified Pscore model gave least bias.

Implementation

Various matching options can be implemented in the GenMatch software (Sekhon, 2011). For example, matching can be performed with or without replacement, with calipers, 1:1 or 1:n, with or without ties. Software and further details can be found at the following web page: `http://sekhon.berkeley.edu/matching/`.

Appendix B

Pscore model in the motivating example

Pscore models for assignment to the DrotAA versus control group were estimated using logistic regression (Rowan et al., 2008). The linear predictor (μ) for the Pscore was:

$$\begin{aligned}\mu &= \beta_0 + \beta_1 \texttt{sex} + \beta_2 \texttt{age} + \beta_3 \texttt{IMscore} + \beta_4 \texttt{icubeds} + \beta_5 \texttt{mva} + \\ &\quad \beta_6 \texttt{orgdys} + \beta_j X_j + \beta_k X_k + \beta_l X_l + \beta_m X_m + \\ &\quad \beta_n X_n + s(\texttt{age}) + s(\texttt{IMscore}),\end{aligned}$$

where `icubeds` indicates the number of beds, `mva` mechanical ventilation, `orgdys` the number of organ failures, X_j, X_k, X_l, X_m, and X_n are vectors of categorical variables for: different types of hospital, sources of admission, serious conditions in the past medical history, types of organ failure, and diagnostic categories. Nonlinearities in the continuous covariates were considered by fitting restricted cubic splines; the terms $s(\texttt{age})$ and $s(\texttt{IMscore})$ represent splines of degree three for `age` and `IMscore`. For the subgroup specific Pscores the same functional form was assumed across subgroups, as is common practice, but separate models were estimated for each subgroup.

The set of variables included in the Pscore or matching algorithm can differ from those for whom balance is presented. The full set of baseline covariates were included in the Pscore and matching algorithms in order to balance major confounders. The choice of variables judged major confounders drew on a previous study which suggested that the most important baseline covariates to balance were IMprob, IMscore, age and the proportion of patients ventilated (Sadique et al, 2011). Here, IMprob is the baseline probability of death (IMprob) which is a function of 20 physiological variables. A second set of variables included in the Pscore were judged potentially weak confounders and of *low priority* to balance (serious conditions in medical history, number of organ failures, types of organ failure). A third set of variables were included in the Pscore and matching algorithm in order to balance the other covariates listed, but were not anticipated to be important confounders (sex, icubeds, hospital type, source of admission, diagnostic category).

Data generating process for simulation study

In scenarios 1 and 3, X_1 and X_2 were generated from the following bivariate normal distribution:

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \overset{iid}{\sim} \mathscr{N}\left(\begin{bmatrix} 2 \\ 4 \end{bmatrix}, \begin{bmatrix} 2 & 0.2 \\ 0.2 & 2 \end{bmatrix}\right),$$

whereas, in scenario 2 from:

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \overset{iid}{\sim} \mathscr{N}\left(\begin{bmatrix} 3 \\ 4 \end{bmatrix}, \begin{bmatrix} 2 & 0.2 \\ 0.2 & 1.5 \end{bmatrix}\right).$$

X_3 was generated from a Bernoulli distribution:

$$X_3 \sim \text{Bern}\begin{cases} 0.6 & \text{for } X_1 > 2 \\ 0.4 & \text{for } X_1 \leq 2 \end{cases},$$

and X_4 was generated from a normal distribution $X_4 \sim \mathscr{N}(3,1)$.

R code for the simulation study

The following code was used to conduct simulations and analyze results for scenario 1a. For all the remaining scenarios the code was modified accordingly. The dataset was generated using the commands:

```
Sigma<-matrix(c(1,0.2,0.2,1),2,2)
X12<-mvrnorm(n,c(2, 4), Sigma)
X1<-X12[,1]
X2<-X12[,2]
X3<-rbern(n,0.5+ifelse(X1>2,0.1,-0.1))
psc_logit<-log(0.2)+(0.1*X1)+(0.2*X2)+(0.3*X3)+(0.2*X1*X3)-
(0.2*X2*X3)
psc<-inv.logit(psc_logit)
tx<-rbern(n,psc)
Y_logit<- -25+(log(2)*tx)+(10*X1)+(0.5*X2)+(0.2*X3)-(log(1.5)*
X3*tx)
Y<-rbern(n,inv.logit(Y_logit))
dataset<-as.data.frame(cbind(X1,X2,X3,Y,tx))
dataset.X3 <- dataset[dataset$X3==1,]
dataset.noX3 <- dataset[dataset$X3==0,] }
```

where `dataset` is the whole sample, and `dataset.X3` and `dataset.noX3` are the sub-samples for the two subgroups, $X_3 = 1$ and $X_3 = 0$. The two Pscore models were fitted separately:

```
pmodel.X3<-glm(tx~X1+X2+X3,family=binomial,data=dataset.X3)
pmodel.noX3<-glm(tx~X1+X2+X3,family=binomial,data=dataset.noX3)
```

and the linear predictors, `pscore.lin.X3` and `pscore.lin.noX3`, and Pscore weights `pscorwght.X3` and `pscorwght.noX3`, calculated. Pscore matching was performed separately for the two subgroups using the commands:

```
mtchout.Y.X3<-Match(Tr=tx,X=cbind(pscore.lin.X3),exact=c(FALSE),
estimand="ATE")
mtchout.Y.noX3<-Match(Tr=tx,X=cbind(pscore.lin.noX3),
exact=c(FALSE),estimand="ATE")
```

GenMatch was performed using the commands:

```
genmtchout.X3<-GenMatch(Tr=tx,X=cbind(pscore.lin.X3,X1,X2),
estimand="ATE", fit.func = my.fitfunc_sdiff,
starting.values=c(10000,0,0),
exact=c(FALSE,FALSE,FALSE),pop.size=gpop)
gmtchout.Y.X3<-Match(Tr=tx,X=cbind(pscore.lin.X3,X1,X2),
exact=c(FALSE,FALSE,FALSE),
Weight.matrix=genmtchout.X3,estimand="ATE")
```

Appendix C

Additional tables and figures

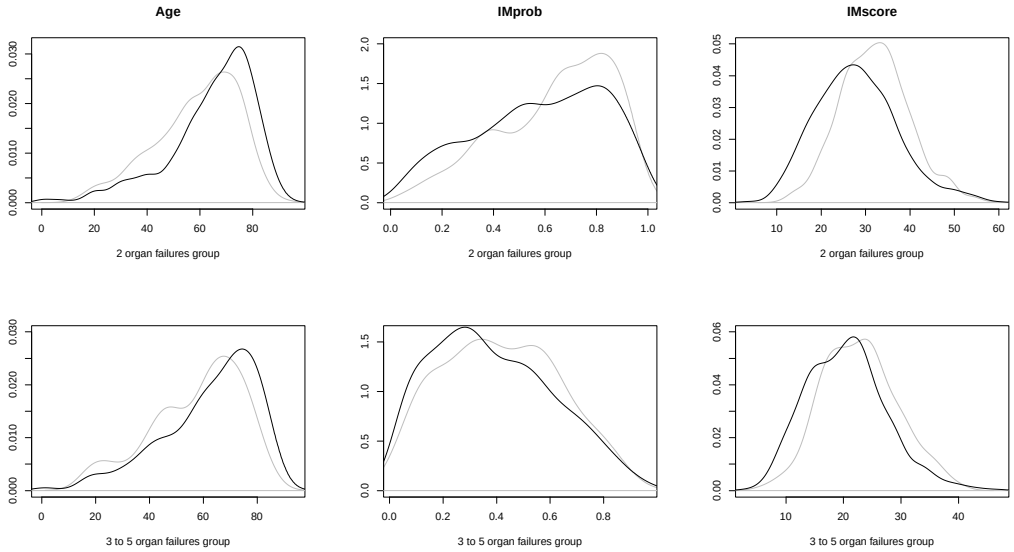

Figure 3: Overlap for the baseline covariates in the motivating example. Density functions reported for age, IMprob and IMscore among treated (grey line) and control (black line) observations by subgroup (2 organ failures; 3 to 5 organ failures).

Table 8: Covariate balance when using i) an overall Pscore or GenMatch algorithm and ii) a subgroup-specific Pscore or GenMatch algorithm: *low priority* variables. Results reported are weighted standardized differences (%)[a].

		Overall Pscore or GenMatch		Subgroup-specific Pscore or GenMatch	
Covariate	Method	2 organ failures group	3 to 5 organ failures group	2 organ failures group	3 to 5 organ failures group
Medical history:					
Cardiovascular					
	Pscore matching	2.99	6.71	25.55	7.05
	GenMatch	18.49	9.06	0.00	7.40
	IPTW	30.75	6.40	11.51	5.08
Respiratory					
	Pscore matching	5.11	6.64	5.17	1.53
	GenMatch	1.32	3.86	5.09	4.78
	IPTW	1.80	2.56	0.94	1.30
Renal					
	Pscore matching	12.14	1.45	11.32	0.97
	GenMatch	16.45	3.16	12.74	2.66
	IPTW	5.43	0.17	2.05	2.45
Liver					
	Pscore matching	33.95	2.54	0.27	0.17
	GenMatch	2.51	0.08	0.00	2.02
	IPTW	16.29	13.04	5.34	29.03
Immunosuppressed					
	Pscore matching	27.86	8.35	3.12	3.93
	GenMatch	5.27	13.45	5.82	12.23
	IPTW	9.06	6.80	2.78	4.33

[a] Note absolute differences are reported as percentages.

Table 9: Covariate balance when using i) an overall Pscore or GenMatch algorithm and ii) a subgroup-specific Pscore or GenMatch algorithm: *low priority* variables. Results reported are weighted standardized differences (%)[a].

		Overall Pscore or GenMatch		Subgroup-specific Pscore or GenMatch	
Covariate	Method	2 organ failures group	3 to 5 organ failures group	2 organ failures group	3 to 5 organ failures group
Number of organ system failing (%):					
3					
	Pscore matching	NA	6.42	NA	2.98
	GenMatch	NA	2.50	NA	4.67
	IPTW	NA	2.54	NA	4.60
4					
	Pscore matching	NA	11.18	NA	9.30
	GenMatch	NA	4.50	NA	5.54
	IPTW	NA	7.71	NA	1.98
5					
	Pscore matching	NA	7.79	NA	10.62
	GenMatch	NA	3.49	NA	1.21
	IPTW	NA	8.48	NA	11.55
Organ system failing in first 24 hours (%)[b]					
Card/Resp					
	Pscore matching	7.61	NA	3.62	NA
	GenMatch	3.19	NA	10.99	NA
	IPTW	6.25	NA	3.66	NA
Card/Resp/Acid					
	Pscore matching	NA	7.19	NA	0.68
	GenMatch	NA	0.45	NA	0.00
	IPTW	NA	5.09	NA	7.11
Card/Resp/ Renal/Acid					
	Pscore matching	NA	9.69	NA	6.21
	GenMatch	NA	5.47	NA	4.38
	IPTW	NA	6.24	NA	1.27

[a] Note absolute differences are reported as percentages. NA: not applicable for a given subgroup.
[b] Results for the most prevalent type for each number of organ failures are presented only (Sadique et al., 2011).

Table 10: Covariate balance of *high priority* variables when using a GenMatch algorithm that optimizes t-tests and KS tests. Results reported are weighted D-statistics [a].

		Subgroup specific Pscore or GenMatch	
Covariate	Method	2 organ failures group	3 to 5 organ failures group
Age			
	Pscore matching	0.06	0.03
	GenMatch	0.05	0.02
	IPTW	0.08	0.19
IMprob			
	Pscore matching	0.15	0.05
	GenMatch	0.05	0.03
	IPTW	0.17	0.06
IMscore			
	Pscore matching	0.10	0.04
	GenMatch	0.05	0.02
	IPTW	0.19	0.15
% Ventilated			
	Pscore matching	0.05	0.01
	GenMatch	0.01	0.00
	IPTW	0.l8	0.15

[a] Note: D-statistics are reported from the weighted version of the KS tests, where weights are frequency weights from matching and IPTW weights. Code available upon request.

Table 11: Covariate balance of *low priority* variables when using a GenMatch algorithm that optimizes t-tests and KS tests. Results reported are weighted D-statistics [a].

		Subgroup specific Pscore or GenMatch	
Covariate	Method	2 organ failures group	3 to 5 organ failures group
Medical history:			
Cardiovascular			
	Pscore matching	0.08	0.01
	GenMatch	0.00	0.01
	IPTW	0.00	0.01
Respiratory			
	Pscore matching	0.00	0.01
	GenMatch	0.01	0.01
	IPTW	0.01	0.03
Renal			
	Pscore matching	0.03	0.00
	GenMatch	0.01	0.01
	IPTW	0.02	0.00
Liver			
	Pscore matching	0.00	0.00
	GenMatch	0.00	0.00
	IPTW	0.00	0.01
Immunosuppressed			
	Pscore matching	0.01	0.02
	GenMatch	0.02	0.04
	IPTW	0.01	0.06

[a] Note: D-statistics are reported from the weighted version of the Kolmogorov-Smirnov tests, where weights are frequency weights from matching and IPTW weights. Code available upon request.

Table 12: Covariate balance of *low priority* variables when using a GenMatch algorithm that optimizes t-tests and KS tests. Results reported are weighted D-statistics [a].

		Subgroup specific Pscore or GenMatch	
Covariate	Method	2 organ failures group	3 to 5 organ failures group
Number of organ system failing (%):			
3			
	Pscore matching	NA	0.02
	GenMatch	NA	0.01
	IPTW	NA	0.07
4			
	Pscore matching	NA	0.05
	GenMatch	NA	0.02
	IPTW	NA	0.05
5			
	Pscore matching	NA	0.04
	GenMatch	NA	0.01
	IPTW	NA	0.01
Organ system failing in first 24 hours (%):			
Card/Resp			
	Pscore matching	3.27	NA
	GenMatch	1.38	NA
	IPTW	2.72	NA
Card/Resp/Acid			
	Pscore matching	NA	0.00
	GenMatch	NA	0.02
	IPTW	NA	0.01
Card/Resp/ Renal/Acid			
	Pscore matching	NA	0.03
	GenMatch	NA	0.01
	IPTW	NA	0.05

[a] Note: D-statistics are reported from the weighted version of the Kolmogorov-Smirnov tests, where weights are frequency weights from matching and IPTW weights. Code available upon request. NA: not applicable for a given subgroup.

Table 13: Motivating example: Covariate balance following IPTW with different levels of weight truncation. Results are reported as weighted standardized differences (%).

Covariate	Percentile	2 organ failures group	3 to 5 organ failures group
Age			
	0,100	9.06	7.24
	1,99	9.13	2.72
	5,95	14.83	8.33
	10,90	20.33	13.29
	25,75	26.19	24.52
IMprob			
	0,100	5.58	9.86
	1,99	7.49	0.27
	5,95	8.17	4.88
	10,90	9.04	7.75
	25,75	10.47	14.01
IMscore			
	0,100	3.88	12.41
	1,99	9.97	4.93
	5,95	16.31	12.89
	10,90	21.66	18.45
	25,75	27.92	30.61
% Ventilated			
	0,100	5.99	13.19
	1,99	13.39	5.54
	5,95	20.23	10.55
	10,90	24.89	14.60
	25,75	30.74	24.44

Table 14: Effectiveness of DrotAA versus controls for each subgroup with: (i) an overall Pscore or GenMatch algorithm and (ii) a subgroup-specific Pscore or GenMatch algorithm. Results are reported as RRs (95% CIs[a]) of hospital mortality.

	Pscore matching	GenMatch	IPTW
i) Overall Pscore or GenMatch			
2 organ failures subgroup	$1.38\ (1.25, 1.50)$	$1.32\ (1.16, 1.46)$	$1.41\ (1.02, 1.81)$
3 to 5 organ failures subgroup	$0.84\ (0.80, 0.88)$	$0.81\ (0.76, 0.86)$	$0.85\ (0.75, 0.95)$
ii) Subgroup-specific Pscore or GenMatch			
2 organ failures subgroup	$1.44\ (1.29, 1.59)$	$1.30\ (1.14, 1.45)$	$1.29\ (1.00, 1.59)$
3 to 5 organ failures subgroup	$0.81\ (0.77, 0.85)$	$0.79\ (0.74, 0.84)$	$0.90\ (0.73, 1.07)$

[a] CIs were calculated using the nonparametric bootstrap.

Table 15: Motivating example: Estimates (CIs[a]) of the coefficients used to obtain the ORs for Table 3 in the main document. The results were obtained by fitting a logistic model with treatment, subgroup, and treatment by subgroup interaction terms as independent covariates. Results are reported for i) an overall Pscore or GenMatch algorithm and ii) a subgroup-specific Pscore or GenMatch algorithm.

	Pscore matching	GenMatch	IPTW
i) Overall Pscore or GenMatch			
Treatment	$0.57\ (0.41, 0.73)$	$0.47\ (0.27, 0.66)$	$0.60\ (0.07, 1.15)$
Interaction	$-0.93\ (-1.11, -0.75)$	$-0.94\ (-1.17, -0.70)$	$-0.95\ (-1.57, -0.36)$
ii) Subgroup-specific Pscore or GenMatch			
Treatment	$0.64\ (0.46, 0.82)$	$0.44\ (0.25, 0.65)$	$0.44\ (0.05, 0.82)$
Interaction	$-1.09\ (-1.18, -0.99)$	$-0.96\ (-1.20, -0.73)$	$-0.69\ (-1.09, -0.26)$

[a] CIs were calculated using the nonparametric bootstrap.

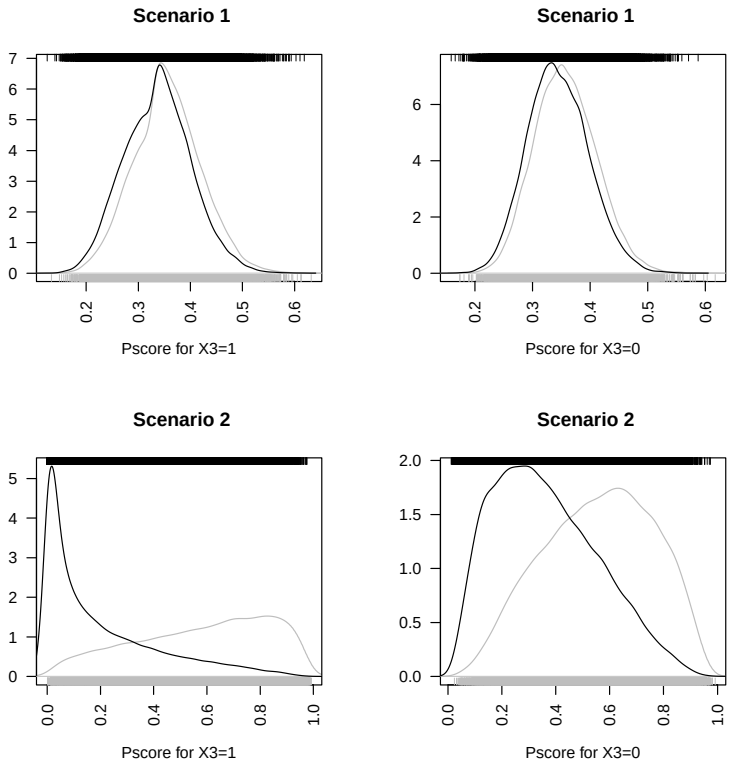

Figure 4: Simulation scenarios with good (Scenario 1) and poor overlap (Scenario 2). Densities of true Pscores using data from a typical sample ($n = 1,000,000$) for treated (grey line) and control (black line). The rug plots, at the top and bottom of each graph, shows the corresponding values of the Pscore.

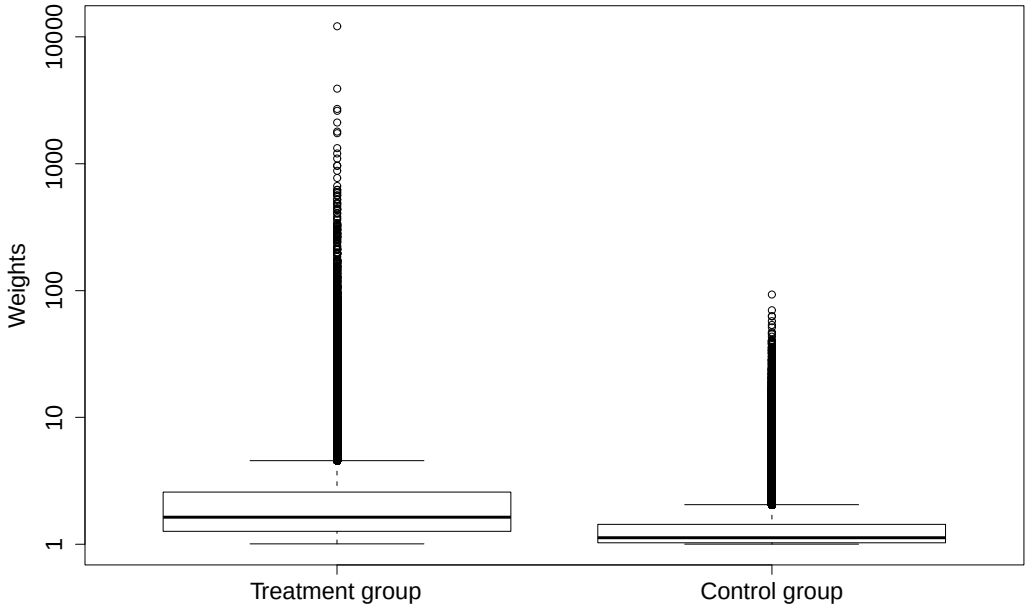

Figure 5: Simulation scenario with poor overlap (2a). Distribution of weights for IPTW for treatment and control observations in the $X_3 = 1$ subgroup, generated for a typical sample ($n = 1,000,000$).

Table 16: Simulation scenario with poor overlap and a misspecified Pscore (2b). Covariate balance following weight truncation. Results are weighted standardized differences (%) reported as averages over the 1,000 replications.

	Subgroup $X_3 = 0$		Subgroup $X_3 = 1$	
Truncation	X_1	X_2	X_1	X_2
0,100	2.59	3.41	16.43	8.88
1,99	2.37	3.75	37.39	11.95
5,95	7.67	15.62	61.86	21.78
10,90	13.16	27.13	75.56	27.43
25,75	24.42	50.90	99.83	37.36

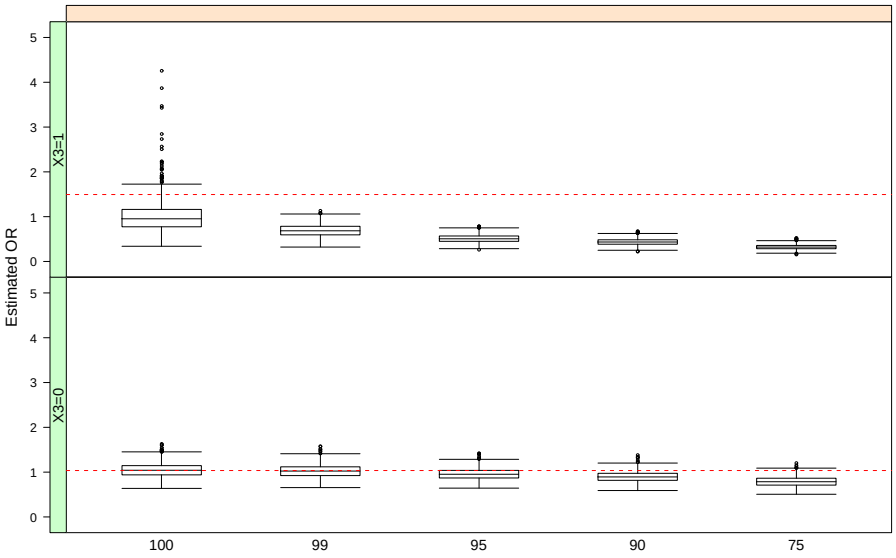

Figure 6: Boxplots of the ORs for IPTW with weight truncation, for the simulation scenario 2b. 100 corresponds to no truncation, 99 corresponds to the case where weights are truncated at the first and 99th percentiles. Results are across 1,000 replications.

References

[1] Abadie, A., Drukker, D., Herr, J.L., and Imbens, G.W. (2001): Implementing matching estimators for average treatment effects in Stata. *The Stata Journal*. 1.

[2] Abadie, A., and Imbens, G.W. (2006): Large sample properties of matching estimators for average treatment effects. *Econometrica*. 74: 235–267.

[3] Abadie, A., and Imbens, G.W. (2009): Matching on the estimated propensity score. *National Bureau of Economic Research*. Available at http://www.hks.harvard.edu/fs/aabadie/research.html

[4] Abraham., E., Laterre, P.F., Garg, R., and *et al* (2005): Administration of Drotrecogin alfa (activated) in early stage severe sepsis (ADDRESS) study group: Drotrecogin alfa (activated) for adults with severe sepsis and a low risk of death. *New England Journal of Medicine*. 353: 1332–1341.

[5] Austin, P.C., Grootendorst, P., and Anderson, G.M. (2007): A comparison of the ability of different propensity score models to balance measured variables between treated and untreated subjects: a Monte Carlo study. *Statistics in Medicine*. 26: 734–753.

[6] Austin, P.C. (2007): The performance of different propensity score methods for estimating marginal odds ratios. *Statistics in Medicine*. 26: 3078–3094.

[7] Austin, P.C. (2008a): A critical appraisal of propensity-score matching in the medical literature between 1996 and 2003. *Statistics in Medicine*. 27: 2037–2049.

[8] Austin, P.C. (2008b): The performance of different propensity-score methods for estimating relative risks. *Journal of Clinical Epidemiology*. 61: 537–545.

[9] Austin, P.C. (2009a): The relative ability of different propensity score methods to balance measured covariates between treated and untreated subjects in observational studies. *Medical Decision Making*. 29: 661–677.

[10] Austin, P.C. (2009b): Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. *Statistics in Medicine*. 28: 3083–3107.

[11] Austin, P.C., and Laupacis, A. (2011): A tutorial on methods to estimating clinically and policy-meaningful measures of treatment effects in prospective observational studies: a review *The International Journal of Biostatistics*. 7.

[12] Bang, H., and Robins, J. (2005). Doubly robust estimation in missing data and causal inference models. *Biometrika*. 61: 692–972.

[13] Bernard, G.R., Vincent, J.L., and Laterre, P.F., and *et al* (2001): Recombinant human protein C worldwide evaluation in severe sepsis (PROWESS) study group: efficacy and safety of recombinant human activated protein C for severe sepsis. *New England Journal of Medicine*. 344: 699–709.

[14] Brookhart, M.A., Schneeweiss, S., Rothman, K.J., Glynn, R.J., Avorn, J., and Stürmer, T. (2006): Variable selection for propensity score models. *American Journal of Epidemiology*. 163: 1149–1156.

[15] Cao, W., Tsiatis, A.A., and Davidian, M. (2009): Improving efficiency and robustness of the doubly robust estimator for a population mean with incomplete data. *Biometrika*. 93: 723–734.

[16] Cole, S.R., and Hernán, M.A. (2008): Constructing inverse probability weights for marginal structural models. *American Journal of Epidemiology*. 168: 656–664.

[17] Diamond, A., and Sekhon, J. (2008): Genetic matching for estimating causal effects: a general multivariate matching method for achieving balance in observational studies. *Working Paper*.

[18] Diamond, A., Sekhon, J. (2012): Genetic matching for estimating causal effects: a general multivariate matching method for achieving balance in observational studies. *Review of Economics and Statistics*. In Press. Available at http://sekhon.berkeley.edu/matching/.

[19] Drake, C. (1993): Effects of misspecification of the propensity score on estimators of treatment effect. *Biometrics*. 49: 1231–1236.

[20] Ertefaie, A., and Stephens, D.A. (2010): Comparing approaches to causal inference for longitudinal data: Inverse probability weighting versus propensity scores. *The International Journal of Biostatistics*. 6.

[21] Gilligan, M.J., and Sergenti, E.J. (2008): Evaluating UN peacekeeping with matching to improve causal inference. *Quarterly Journal of Political Science*. 3: 89–122.

[22] Gordon, S., and Huber, G. (2007): The effect of electoral competitiveness on incumbent behavior. *Quarterly Journal of Political Science*. 2: 107–138.

[23] Greenand, S.,Robins, M.R., and Pearl, J. (1999): Confounding and collapsibility in causal inference. *Statistical Science*. 14.

[24] Grieve, R., Sekhon, J., Hu, T., and Bloom, J. (2008): Evaluating health care programs by combining cost with quality of life measures: a case study comparing capitation and fee for service. *Health Services Research*. 43: 1204–1222.

[25] Hasford, J., Bramlage, P., Koch, G., Lehmacher, W., Einhäupl, K., and Rothwell, P.M. (2010): Inconsistent trial assessments by the National Institute for Health and Clinical Excellence and IQWiG: standards for the performance and interpretation of subgroup analyses are needed. *Journal of Clinical Epidemiology*. 63: 1298–1304.

[26] Heinrich, C.J. (2008): Demand and supply-side determinants of conditional cash transfer program effectiveness. *World Development*. 35: 121–143.

[27] Hernán, M.A.B., Brumback, B., and Robins, J.M. (2000): Marginal structural models to estimate the causal effect of zidovudine on the survival of HIV-positive men. *Epidemiology*. 11: 561–570.

[28] Herron, M.C., and Wand, J. (2007): Assessing partisan bias in voting technology: The case of the 2004 new Hampshire recount. *Election Studies*. 26: 247–261.

[29] Hirano, K., Imbens, G., and Ridder, G. (2003): Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*. 71: 1161–1189.

[30] Imai, K., King, G., and Stuart, E.A. (2008): Misunderstandings between experimentalists and observationalists about causal inference. *Journal of Royal Statistical Society: Series A*. 171: 481–502.

[31] Kang, J.D., and Schafer, J.L. (2007): Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical Science*. 22: 523–539.

[32] Korkeamaki, O., and Uusitalo, R. (2009): Employment and wage effects of a payroll-tax cut evidence from a regional experiment. *International Tax and Public Finance*. 16: 753–772.

[33] Kurth, T., Walker, A.M., Glynn, R.J., and *et al* (2006): Results of multivariable logistic regression, propensity matching, propensity adjustment, and propensity-based weighting under conditions of nonuniform effect. *American Journal of Epidemiology*. 163: 262–270.

[34] Kuss, O., Legler, T., and Börgermann, J. (2011): Treatments effects from randomized trials and propensity score analyses were similar in similar populations in an example from cardiac surgery. *Journal of Clinical Epidemiology*. 64: 1076–1084.

[35] Lefebvre, G., and Gustafson, P. (2010): Impact of outcome model misspecification on regression and doubly-robust inverse probability weighting to estimate causal effect. *The International Journal of Biostatistics*. 6.

[36] Lenz, G.S., and Ladd, J.M. (2009): Exploiting a rare communication shift to document the persuasive power of the news media. *American Journal of Political Science*. 53: 394–410.

[37] Lunceford, J.K., and Davidian, M. (2004): Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. *Statistics in Medicine*. 23: 2937–2960.

[38] Mattei, A. (2009): Estimating and using propensity score in presence of missing background data: an application to assess the impact of childbearing on wellbeing. *Statistical Methods & Applications*. 18: 257-273.

[39] Mebane, Jr W.R., and Sekhon, J.S. (2011): Genetic optimization using derivatives: The rgenoud package for R. *Journal of Statistical Software*. 42: 1–26.

[40] Moodie, E.E.M., and Stephens, D.A. (2010): Special issue on causal inference. *The International Journal of Biostatistics*. 6.

[41] Normand, S.L.T., Landrum, M.B., Guadagnoli, E. and *et al* (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.

[42] Pearl, J. (1995): Causal diagrams for empirical research. *Biometrika*. 82: 669–710.

[43] Petersen, M.L., Porter, K., Gruber, S., Wang, Y., and van der Laan, M.J.V.D. (2012): Diagnosing and responding to violations in the positivity assumption. *Statistical Methods in Medical Research*. 21: 31–54.

[44] Porter, K.E., Gruber, S., van der Laan, M.J., and Sekhon, J.S. (2011). The relative performance of targeted maximum likelihood estimators. *The International Journal of Biostatistics*. 6.

[45] Qu, Y., and I, Lipkovich. (2009): Propensity score estimation with missing values using a mutiple imputation missingness pattern (MIMP) approach. *Statistics in Medicine*. 28: 1402-1414.

[46] Ramsahai, R., Grieve, R., and Sekhon, J.S. (2011): Extending iterative matching methods: An approach to improving covariate balance that allows prioritisation. *Health Service Outcomes Research Methodology*. 11: 95–114

[47] Robins, J.M., Rotnitzky, A., and Zhao, L.P. (1994): Estimation of regression coefficients when some regressors are not always observed. *Journal of American Statistical Association*. 89: 846–866.

[48] Robins, J.M., Hernán, M.A., and Brumback, B. (2000): Marginal structural models and causal inference in epidemiology. *Epidemiology*. 11: 550–560.

[49] Robins, J.M., Sued, M., Lei-Gomez, Q., and Rotnitzky, A. (2007): Comment: performance of double-robust estimators when inverse probability weights are highly variable. *Statistical Science*. 22: 544–559.

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

[51] Rosenbaum, P., and Rubin, D. (1984): Reducing bias in observational studies using subclassification on the propensity score. *Journal of American Statistical Association*. 79: 516–524.

[52] Rosenbaum, P.R., and Rubin, D.B. (1985): Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *American Statistician*. 39: 33–38.

[53] Rowan, K.M., Welch, C.A., North, E., and Harrison, D.A. (2008): Drotrecogin alfa (activated): real-life use and outcomes for the UK. *Critical Care*. 12: R58.

[54] Rubin, D.B., and Thomas, N. (1992): Affinely invariant matching methods with ellipsoidal distributions. *Annals of Statistics*. 20: 1079–1093.

[55] Rubin, D.B. (1997): Estimating causal effects from large data sets using propensity scores. *Annals of Internal Medicine*. 127: 757–763.

[56] Rubin, D.B. (2008): For objective causal inference, design trumps analysis. *Annals of Applied Statistics*. 2: 808–840.

[57] Sadique, M.Z., Grieve, R., Harrison, D.A., Cuthbertson, B.H., Rowan, K.M. (2011): Is Drotrecogin alfa (activated) for adults with severe sepsis, cost-effective in routine clinical practice?. *Critical Care*. 15: 1–15.

[58] Scharfstein, D.O., Rotnitzky, A., and Robins, J.M. (1999): Rejoinder to adjusting for nonignorable drop-out using semiparametric nonresponse models. *Journal of American Statistical Association*. 94: 1135–1146.

[59] Schisterman, E.F., Cole, S.R., and Platt, R.W. (2007): Overadjustment bias and unnecessary adjustment in epidemiologic studies. *Epidemiology*. 20: 488–495.

[60] Sekhon, J.S., and Mebane, Jr W.R. (1998): Genetic optimization using derivatives: Theory and application to nonlinear models. *Political Analysis*. 7: 189–203.

[61] Sekhon, J. (2011): Matching: multivariate and propensity score matching with automated balance search. *Journal of Statistical Software*. In Press. Computer program available at http://sekhon.berkeley.edu/matching/.

[62] Sekhon, J., and Grieve, R. (2011): A matching method for improving covariate balance in cost-effectiveness analyses. *Health Economics*. 21: 695-714

[63] Shah, B.R., Laupacis, A., Hux, J.E., and Austin, P.C. (2005): Propensity score methods give similar results to traditional regression modelling in observational studies: a systematic review. *Journal of Clinical Epidemiology*. 58: 550–559.

[64] Stuart, E.A. (2010): Matching methods for causal inference: a review and a look forward. *Statistical Science*. 25: 1–21.

[65] Stuart, E.A., and Green, K.M. (2008): Using full matching to estimate causal effects in nonexperimental studies: examining the relationship between adolescent marijuana use and adult outcomes. *Developmental Psychology*. 44: 395–406.

[66] Stürmer, T., Joshi, M., Glynn, R.J., Avorn, J., Rothman, K.J., and 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.

[67] van der Laan, M.J. (2010a): Targeted maximum likelihood based causal inference: Part I. *International Journal of Biostatisitics*. 6.

[68] van der Laan, M.J. (2010b): Targeted maximum likelihood based causal inference: Part II. *International Journal of Biostatisitics*. 6.

[69] van der Laan, M.J., and Gruber, S. (2010): Collaborative double robust targeted maximum likelihood estimation. *International Journal of Biostatisitics*. 6.

[70] Woo, M., Reiter, P.J., and Karr, F.K. (2008): Estimation of propensity scores using generalized additive models. *Statistics in Medicine*. 27: 3805–3816.

[71] Zhao, Z. (2008): Sensitivity of propensity score methods to the specifications. *Economics Letters*. 98: 309–319.