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

Combining propensity score-based stratification and weighting to improve causal inference in the evaluation of health care interventions

Ariel Linden DrPH[1,2]

[1]President, Linden Consulting Group, Ann Arbor, MI, USA
[2]Adjunct Associate Professor, Department of Health Management & Policy, School of Public Health, University of Michigan, Ann Arbor, MI, USA

Keywords
marginal mean weighting through stratification, observational studies, propensity score, stratification, weighting

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

Authors' Note: All of the proposed methods in this paper can be implemented via companion software for Stata which was written by the author and is available on request.

Accepted for publication: 15 July 2014

doi:10.1111/jep.12254

Abstract

When a randomized controlled trial is not feasible, a key strategy in observational studies is to ensure that intervention and control groups are comparable on observed characteristics and assume that the remaining unmeasured characteristics will not bias the results. In the past few years, propensity score-based techniques such as matching, stratification and weighting have become increasingly popular for evaluating health care interventions. Recently, marginal mean weighting through stratification (MMWS) has been introduced as a flexible pre-processing approach that combines the salient features of propensity score stratification and weighting to remove imbalances of pre-intervention characteristics between two or more groups under study. The weight is then used within the appropriate outcome model to provide unbiased estimates of treatment effects. In this paper, the MMWS technique is introduced by illustrating its implementation in three typical experimental conditions: a binary treatment (treatment versus control), an ordinal level treatment (varying doses) and nominal treatments (multiple independent arms). These methods are demonstrated in the context of health care evaluations by examining the pre-post difference in hospitalizations following the implementation of a disease management program for patients with congestive heart failure. Because of the flexibility and wide application of MMWS, it should be considered as an alternative procedure for use with observational data to evaluate the effectiveness of health care interventions.

Introduction

Conducting a randomized controlled trial (RCT) to evaluate the effectiveness of health-related programs and other interventions is often not feasible due to logistical, practical or ethical reasons. In such situations, program evaluators attempt to emulate the randomization process in observational data by ensuring that all groups under study are comparable on observed pre-intervention characteristics. However, unlike in an RCT, an assumption is required that unmeasured characteristics do not bias the results [1]. When pre-intervention differences between groups are found, conventional regression modeling remains the most common adjustment approach, even though there is sufficient evidence that these methods may produce biased results [2,3].

In recent years, adjustment techniques based on the propensity score have become increasingly popular in health care evaluations [4–7]. The propensity score reflects the probability of assignment to the treatment group conditional on observed covariates [8]. Propensity scores are generally derived from a logistic regression model that reduces each participant's set of covariates to a single score. Conditional on a well-constructed propensity score, pretreatment covariates will be independent of group assignment and will be distributed similarly across study groups. When correlation of covariates and treatment assignment is removed, the covariates will not confound estimated treatment effects [8]. Once the propensity score has been estimated in a given dataset, a data 'preprocessing' procedure is performed to create comparability between study groups and typically involves matching, stratification or weighting. It is referred to as pre-processing because it is performed before the final treatment effect is estimated, thus replicating the RCT by separating the study design stage from the outcomes analysis [5].

There are several different matching algorithms currently in use to match treated to non-treated individuals on their propensity score, such as pairwise matching (also called one-to-one matching), 1:*k* matching [9], matching using propensity score categories

[10], matching based on the Mahalanobis distance [11] and kernel density matching [12] (see Caliendo & Kopeinig [13] for a comprehensive discussion on propensity score matching).

Stratification (also referred to as subclassification) is another propensity score adjustment approach. Here, the entire range of propensity scores in the dataset is divided into strata (by partitioning the propensity score into a number of quantiles), and then treated and non-treated groups are arranged within each stratum accordingly. This approach allows the evaluator to analyse outcomes between groups within each stratum as well as to observe overall differences between groups across all strata [14]. It has been shown that stratification of the propensity score into five quantiles can remove over 90% of the initial bias due to the covariates used to create the propensity score [15,16].

Another propensity score-based adjustment procedure involves weighting each individual in the data, conditional on both their propensity score and treatment group assignment. The most commonly used weighting scheme is the inverse probability of treatment weights (IPTW) [3,17], which is intended to standardize the treatment groups to the population for which treatment is intended. Participants receive a weight equal to the inverse of the estimated propensity score (1/propensity score) and non-participants receive a weight equal to the inverse of 1 minus the estimated propensity score (1/(1 – propensity score)). Once the weights are constructed, they can then be used within the appropriate regression model framework. A recent extension of the IPTW approach is the 'doubly robust' estimator [18,19], which utilizes IPTW and covariates within the same outcome model. An estimator is considered doubly robust as long as either model (propensity score or outcomes) is correctly specified. Therefore, an evaluator is given two chances, instead of only one, to make a valid inference about the effects of the treatment.

Recently, an approach has been introduced that combines elements of both propensity score stratification and IPTW [20–22]. In general, this first entails stratifying the analytic sample into quantiles of the propensity score, and then generating a weight for each individual based on their corresponding stratum and treatment assignment. The stratification reduces bias in the observed covariates used to create the propensity score, and the weighting standardizes each treatment group to the target population.

This approach, named 'marginal mean weighting through stratification' (MMWS) [21,22], can handle a broad array of experimental conditions that researchers will likely encounter in evaluating health care interventions such as: binary treatments (one treatment and one control group), ordinal treatments (various dose levels of a treatment) and nominal treatments (multiple independent treatments). Once generated, the MMWS can then be used within the appropriate outcome model to estimate unbiased treatment effects.

This paper introduces MMWS as an alternative propensity score-based approach for providing unbiased treatment effect estimates in non-randomized health care interventions. Its application is illustrated using data from a disease management (DM) program for patients with congestive heart failure. The paper begins by describing the dataset and is followed by a detailed explanation of the analytic procedure for binary, ordinal and nominal treatments applied to the current data. The final section discusses the strengths and weaknesses of the MMWS.

Data

Our data come from a DM program designed for patients with congestive heart failure and implemented in a large health plan located in the Western United States. Individuals with the condition were called and invited to enrol in the program. Those agreeing to participate received one of the following interventions based on the subjective assignment by a program nurse: (1) periodic telephone calls from a nurse to discuss self-management behaviours; or (2) remote tele-monitoring (RTM) that entailed daily electronic transmission of the participant's disease-related symptoms to a database followed by a call from the nurse if symptoms appeared to indicate the onset of an acute exacerbation. The primary goal of the intervention was to reduce avoidable hospitalizations [23]. We use these data solely to illustrate the MMWS techniques, and our analyses do not represent a definitive assessment of the program's effectiveness. These data were chosen because the intervention can be examined in various ways – as a binary treatment (participants versus non-participants), as an ordinal treatment (varying doses of the telephonic intervention) or as a nominal treatment (comparison between telephonic and RTM interventions).

The retrospectively collected data consist of observations for 1359 program participants who completed a full 12 months of the intervention and 6612 non-participants who were health plan members during the same period but were not exposed to the intervention. Each individual in the dataset has 12 months of pre-intervention data and 12 months of intervention-period data. The primary outcome for all analyses used in this paper is the difference between pre-intervention and intervention-period all-cause hospitalization rates. All analyses were conducted using a software program written by the author for Stata (StataCorp., College Station, TX, USA), which is available upon request.

Application of MMWS to a binary treatment

A binary treatment (in which a treatment group is compared with a control group), is the most prevalent study design in health care. In this framework, a well-matched control group serves as the counterfactual to the treatment group, that is, it provides an estimate of what the treatment group's outcome would have been had it simultaneously not received the treatment. Here, we compare all 1359 program participants to all 6612 non-participants.

Supporting Information Appendix S1 describes the pre-intervention characteristics of the DM program participant and non-participant groups together with their unadjusted standardized differences [24] and *P*-values (presented as measures of covariate balance between groups). If the groups were comparable, standardized differences would be close to zero and *P*-values would be non-significant (>0.05). However, there are several observed baseline variables that are imbalanced (standard differences >0.10 or *P*-values < 0.05), indicating the need to adjust for selection bias.

The first step in implementing MMWS is to estimate the propensity score using logistic regression and save the predicted values for each individual. Here, the binary treatment variable was regressed on all variables presented in Supporting Information Appendix S1, including patient demographic characteristics (age and gender), the Charlson comorbidity index and associated

Table 1 Calculation of the marginal mean weights through stratification (MMWS) for a binary treatment within common support

Stratum	Treated ($n_{z=1,s}$)	MMWS (treated)	Non-treated ($n_{z=0,s}$)	MMWS (non-treated)	n_s
1	121	2.247	1473	0.898	1594
2	196	1.387	1397	0.946	1593
3	236	1.152	1357	0.974	1593
4	355	0.766	1238	1.067	1593
5	451	0.603	1142	1.157	1593
Total	1,359		6607		7966

comorbidities [25], and key measures of health care utilization (prescription filled, office visits, emergency department visits, hospital admissions and hospital days).

Next, the region of common support was identified and individuals outside of this region were flagged. Common support simply means that treated individuals have a corresponding counterfactual. Most commonly, treated individuals at the high end of the propensity score distribution may not have non-treated individuals with corresponding propensity scores, and vice versa. Treatment effect estimates will be biased if individuals do not have counterfactuals as a point of comparison. In the MMWS framework, individuals outside the region of common support receive a weight of zero [22]. In our data, five individuals (all from the control group) were outside the region of support and received weights of zero.

Next, propensity scores for all individuals were stratified into five approximately equal-sized quintiles of 1593 individuals per stratum (see Table 1), as recommended by Rosenbaum & Rubin [16].

Finally, the marginal mean weights for the binary treatment were computed based on the following equation by Hong [22]:

$$\frac{n_s \times \Pr(Z=z)}{n_{z=z,s}} \tag{1}$$

where n_s is the total number of individuals in a given stratum s, $\Pr(Z=z)$ is the probability of assignment to treatment group z, and $n_{z=z,s}$ is the total number of individuals in stratum s that were actually assigned to treatment z. Table 1 displays all the values needed to compute the MMWS for each stratum by group assignment. Using the treatment group in stratum 1 as an example, we replace Equation 1 with the numeric equivalents:

$$\frac{1594 \times \left(\frac{1359}{7966}\right)}{121} = 2.247$$

Similarly, the MMWS weights for the control group in stratum 1 are calculated as follows:

$$\frac{1594 \times \left(\frac{6607}{7966}\right)}{1473} = 0.898$$

After weights were calculated for each treatment group by stratum, each individual in the dataset received the weight corresponding to their stratum and treatment assignment.

Supporting Information Appendix S2 presents the baseline characteristics of the treatment and control groups adjusted with the MMWS weights. As shown, all baseline covariates now have standardized differences closer to zero than when unweighted, and all P-values are substantially higher than the traditional 0.05 cut-off. This lends greater confidence that the weighting approach has controlled for observed confounding that may bias the treatment effect estimates.

Supporting Information Appendix S3 provides a comparison of the unweighted and weighted treatment effect estimates for the differences in pre-to-post hospitalization rates. In the unweighted data, the treatment group experienced a reduction in pre-post admissions of –0.20 admission per person while the control group had a substantially larger reduction in pre-post admission of –0.33 per person. The net difference of 0.13 in favour of the control group was statistically significant ($P < 0.0001$). After adjustment using MMWS, the control group still had a favourable difference in pre-post admissions versus the treatment group, but the difference was no longer statistically significant (difference in differences = 0.08, P-value = 0.056, 95% CI: –0.002, 0.165).

Application of MMWS to ordinal treatments

Interventions that include varying doses of an ordinal level treatment are also pervasive in health care. Examples include drug studies in which different dosages of a medication are compared and health interventions that involve multiple contacts with a health professional. In these cases, the evaluation seeks to assess whether there is a positive (or inverse) relationship between dose and the outcome.

In the multi-dose framework, the comparison group may be either a control group that receives no intervention or the group that receives the lowest treatment dose (in studies where all participants receive at least some level of the intervention). In an observational study, weighting techniques can serve to achieve balance on baseline characteristics between all treatment levels. In the outcomes analysis, either the control group or the lowest dose group (in those studies in which everyone is exposed to some level of the intervention) serves as the counterfactual for the treatment groups at higher dosages.

In the current data, there were 654 participants in the 12-month telephonic intervention. For the purpose of illustrating the MMWS technique, the sample is divided into four distinct groups based on the observed distribution of the call frequency: (1) non-participants; (2) participants who received only one call; (3) participants who received two or three calls; and (4) participants who received four or more calls over the 12-month intervention period. Group 1 serves as the control and is comprised of non-participants who never received any calls.

Table 2 Calculation of the marginal mean weights through stratification (MMWS) for an ordinal treatment within common support

Stratum	Non-participants ($n_{z=0,s}$)	MMWS (no calls)	1 Call ($n_{z=1,s}$)	MMWS (one call)	2–3 Calls ($n_{z=2,s}$)	MMWS (2–3 calls)	4 + Calls ($n_{z=3,s}$)	MMWS (4 + calls)	n_s
1	1181	1.048	77	0.873	62	0.616	46	0.500	1366
2	1210	1.022	72	0.933	53	0.721	30	0.766	1365
3	1235	1.002	78	0.862	29	1.318	24	0.959	1366
4	1281	0.965	51	1.317	25	1.528	8	2.874	1365
5	1278	0.968	58	1.158	22	1.736	7	3.285	1365
Total	6185		336		191		115		6827

Supporting Information Appendix S4 describes the unadjusted pre-intervention characteristics of participants in the three dose levels of the intervention and the non-participant group. As shown, groups were statistically different from each other in 7 of the 11 covariates – once again indicating the need to adjust for selection bias.

For an intervention with ordinal level treatments, the propensity score can be estimated via ordinal logistic regression [26]. Ordinal regression assumes that the proportional odds of being assigned to any given treatment level are the same as any other treatment level, given the covariates used for estimation (this is referred to as the parallel regression assumption). If this assumption holds, any predicted level of treatment may serve as the basis for calculating the propensity score. In the current data, the ordinal treatment variable was regressed on all covariates described earlier in the binary treatment section. Following the recommendation of Lu *et al*. [27], the propensity score was estimated to reflect the probability of being assigned to the lowest treatment dose (non-participation group).

As in the case of the binary treatment, the region of common support was established by ensuring that no individual in any group exceeded the lower or upper common propensity score threshold of any other treatment level. All individuals outside of this common region were flagged and received an MMWS weight of zero. In the current data, a total of 439 individuals were flagged as being outside the region of support, leaving a total of 6827 individuals for the analytic sample (see Table 2).

Next, propensity scores for everyone in the analytic sample were stratified into five approximately equal-sized quintiles of 1365 individuals per stratum and the marginal mean weights were computed based on the same formula as presented in Equation 1.

Table 2 displays all the necessary values used in computing the MMWS for each stratum and treatment assignment. Using the non-treatment group in stratum 1 as an example, we replace Equation 1 with the numeric equivalents:

$$\frac{1366\times\left(\frac{6185}{6827}\right)}{1181}=1.048$$

Similarly, the MMWS weights for the group receiving four or more calls during the intervention year in stratum 2 are calculated as follows:

$$\frac{1365\times\left(\frac{115}{6827}\right)}{30}=0.766$$

After weights were calculated for each treatment group by stratum, each individual in the dataset received the weight corresponding to their stratum and treatment assignment.

Supporting Information Appendix S5 presents the baseline characteristics of the participants in the three dose levels of the intervention compared with the non-participant group adjusted using MMWS weights. As there is no equivalent to the standardized difference metric for multiple treatments, researchers mostly rely on *P*-values as the numeric measure of covariate balance, with the statistical expectation being that up to 5% of the covariates analysed will be statistically significant due to chance alone. As shown, with the exception of age, all covariates in the weighted data have *P*-values that are substantially higher than the traditional 0.05 cut-off. The residual imbalance of age will be further adjusted by adding it as a covariate in the regression model for the outcome analysis [28].

Supporting Information Appendix S6 provides the weighted difference-in-difference estimates of hospitalizations for each of the three treatment levels and non-participants as well as Bonferroni adjusted contrasts between each level versus non-participants. Taken together, both the non-treatment group and the single call per year group showed a decrease in pre-post admissions, but the difference between them was not statistically significant (*P*-value = 0.75). Contrary to expectations, both the 2–3 calls per year group and 4+ calls per year group increased their pre-post admissions to a greater extent than non-participants.

Application of MMWS to nominal treatments

In some health care studies, patients are assigned to one of several independent (nominal) interventions. For example, a drug study may be implemented to compare the efficacy of competing drugs, or a health management program may utilize different modes of communication with patients to determine which approach results in the highest level of patient self-management.

In the nominal treatment framework, no treatment group is considered higher in ranking order than any other (except when compared with the control group) and all treatment groups are expected to be comparable on baseline characteristics. In the outcomes analysis of nominal treatments, all groups serve as counterfactuals to all others with a series of post-estimation contrasts conducted to determine which intervention was the most effective.

Table 3 Calculation of the marginal mean weights through stratification (MMWS) for multiple (nominal) treatments within common support

	Non-participants			Telephonic			RTM		
Stratum	n_{s0}	$n_{z=0,s0}$	MMWS	n_{s1}	$n_{z=1,s1}$	MMWS	n_{s2}	$n_{z=2,s2}$	MMWS
1	1574	1128	1.156	1574	78	1.665	1574	33	4.256
2	1573	1219	1.069	1573	105	1.236	1573	82	1.712
3	1574	1340	0.973	1574	126	1.031	1574	116	1.211
4	1573	1371	0.950	1573	158	0.821	1573	178	0.789
5	1573	1458	0.894	1573	182	0.713	1573	293	0.479
Total	7867	6516		7867	649		7867	702	

RTM, remote tele-monitoring.

For the purpose of illustrating the MMWS technique for nominal treatments, the current sample was divided according to treatment assignment: (1) 6612 non-participants; (2) 654 participants in the telephonic intervention; and (3) 705 participants in the RTM intervention. Supporting Information Appendix S7 presents the unadjusted pre-intervention characteristics of participants in the three study arms. As shown, groups were statistically different from each other in 6 of the 11 covariates.

For an intervention with nominal treatments, propensity scores are estimated using multinomial logistic regression [26]. Using this approach, each individual receives one propensity score corresponding to the probability of assignment to each treatment, conditional on baseline characteristics. Thus, in the current data, three propensity scores were estimated for each individual corresponding to their probability of assignment to non-participation, the telephonic intervention and RTM, respectively.

The region of common support was determined by ensuring that each individual's three propensity scores were within the common bounds of all three propensity score ranges. Individuals outside of this common region were flagged and received an MMWS weight of zero. In the current data, a total of 104 individuals were flagged as being outside the common region of support, leaving 7867 individuals remaining in the analytic sample.

Next, each of the three propensity scores for everyone in the analytic sample was stratified into five approximately equal-sized quintiles of 1573 individuals per stratum. Thus, there were three propensity score strata (corresponding to the three propensity scores) divided into five quintiles each (see Table 3). The marginal mean weights for the nominal case were computed based on the formula by Hong [22]:

$$\frac{n_{s_z} \times \Pr(Z=z)}{n_{z=z,s_z}} \tag{2}$$

where n_{s_z} is the total number of individuals in a given stratum s_z for a corresponding propensity score θ_z, $\Pr(Z = z)$ is the probability of assignment to a given treatment group z, and $n_{z=z,s_z}$ is the total number of individuals in stratum s_z that were actually assigned to treatment z. Table 3 displays all the necessary values used in computing the MMWS for each stratum and group assignment in the nominal treatment case. Taking the RTM group in stratum 1 as an example, we replace Equation 2 with the numeric equivalents:

$$\frac{1547 \times \left(\frac{702}{7867}\right)}{33} = 4.256$$

Similarly, the MMWS weights for the telephonic intervention in stratum 1 are calculated as follows:

$$\frac{1574 \times \left(\frac{649}{7867}\right)}{78} = 1.665$$

After weights were calculated for each treatment group by stratum, each individual in the dataset received the weight corresponding to their stratum and their actual treatment assignment.

Supporting Information Appendix S8 presents the baseline characteristics of participants in the three arms of the study, adjusted with the MMWS weights. As shown, with the exception of hospital days, all covariates in the weighted data have *P*-values that are substantially higher than the traditional 0.05 cut-off. The residual imbalance of hospital days is further adjusted by adding it as a covariate in the regression model for the outcome analysis.

Supporting Information Appendix S9 provides the weighted difference-in-difference estimates of hospitalizations for the three study arms, as well as Bonferroni adjusted contrasts between each arm and the other two arms. Both the non-treatment group and the RTM group showed a comparable decrease in pre-post admissions of –0.30 admission per person, while the telephonic intervention had a much lesser decrease of –0.15 admission per person. The contrasts indicate that there was no statistical difference between RTM versus non-participants, or RTM versus the telephonic intervention. However, the telephonic intervention group did statistically worse than the non-treatment group (0.15 admission per person, 95% CI 0.019, 0.278).

Discussion

The purpose of this paper was to introduce readers to MMWS, a recent addition to the family of propensity score-based techniques. The analytic approach was described in detail for three of the most typical types of experimental designs (the binary treatment, ordinal treatments and nominal treatments) using data from a DM intervention.

Because of its flexibility and wide application, MMWS may hold greater appeal for program evaluators compared with existing approaches, most of which are constrained by the specific study design implemented. For example, while the entire array of propensity scoring techniques (described in the Introduction) can be used to evaluate binary treatments, only weighting approaches can be seriously considered for evaluating interventions with multiple levels or multiple arms [3]. While this does not pose a limitation in evaluating health care interventions per se, an attractive quality of MMWS is the ability to conduct several types of evaluations of the same intervention using a consistent overall approach (as demonstrated by our examples). Moreover, the MMWS approach can be extended to accommodate multiple concurrent treatments, handle multilevel data and assess the effects of moderated treatments [21,22]. A somewhat similar approach has also been proposed to analyse treatment effects in longitudinal data [29].

Beyond its flexibility, MMWS has also been shown to be more accurate compared with IPTW in estimating outcomes. Huang *et al*. [20] used both techniques and found that the IPTW results were much more variable, and in many cases did not agree with the other two methods applied to the data (the stratification approach and hierarchical outcome regression). Similarly, Hong [21] found through a comprehensive set of simulations that MMWS had much lower bias and higher accuracy (lower mean square errors) than IPTW when the propensity score model was misspecified (which is the case with most data used in health care interventions).

While MMWS may be more appealing than other propensity scoring approaches for evaluating health care interventions using non-experimental data, it still carries the same limitation as all other models for causal inference; that is, it assumes that all biases and sources of confounding have been adjusted for in the model – an assumption that cannot be tested outside of a randomized study. Thus, regardless of the specific technique employed, the evaluation is best served by conducting sensitivity analyses to determine the magnitude of unmeasured bias necessary to alter the conclusion that observed outcomes reflect the effect of the intervention [30].

In summary, this paper describes an alternative propensity score-based adjustment procedure that combines stratification and weighting to allow for the estimation of treatment effects. Because of its flexibility and wide application, MMWS should be considered as an alternative procedure for use with observational data when evaluating the effectiveness of health care interventions.

Acknowledgement

I would like to thank Julia Adler-Milstein for her review and helpful comments.

References

1. Rubin, D. B. (2007) The design versus the analysis of observational studies for causal effects: parallels with the design of randomized trials. *Statistics in Medicine*, 26, 20–30.
2. Freedman, D. (1999) From association to causation: some remarks on the history of statistics. *Statistical Science*, 14, 243–258.
3. Robins, J. M., Hernán, M. A. & Brumback, B. (2000) Marginal structural models and causal inference in epidemiology. *Epidemiology (Cambridge, Mass.)*, 11, 550–560.
4. Sturmer, T., Joshi, M., Glynn, R. J., Avorn, J., Rothman, K. J. & Schneeweiss, S. (2006) A review of the application of propensity score methods yielded increasing use, advantages in specific settings, but not substantially different estimates compared with conventional multivariable methods. *Journal of Clinical Epidemiology*, 59, 437–447.
5. Stuart, E. A. (2010) Matching methods for causal inference: a review and a look forward. *Statistical Science*, 25, 1–21.
6. Austin, P. C. (2007) Propensity-score matching in the cardiovascular surgery literature from 2004 to 2006: a systematic review and suggestions for improvement. *Journal of Thoracic Cardiovascular Surgery*, 134, 128–1135.
7. Austin, P. C. (2008) A critical appraisal of propensity-score matching in the medical literature between 1996 and 2003. *Statistics in Medicine*, 27, 2037–2049.
8. Rosenbaum, P. R. & Rubin, D. B. (1983) The central role of propensity score in observational studies for causal effects. *Biometrika*, 70, 41–55.
9. Linden, A. & Samuels, S. J. (2013) Using balance statistics to determine the optimal number of controls in matching studies. *Journal of Evaluation in Clinical Practice*, 19, 968–975.
10. Dehejia, R. H. & Wahba, S. (1999) Causal effects in nonexperimental studies: reevaluating the evaluation of training studies. *Journal of the American Statistical Association*, 94, 1053–1062.
11. Rubin, D. B. (1980) Bias reduction using Mahalanobis metric matching. *Biometrics*, 36, 293–298.
12. Heckman, J., Ichimura, J. & Todd, P. (1997) Matching as an econometric evaluation estimator: evidence from evaluating a job training programme. *Review of Economic Studies*, 64, 605–654.
13. Caliendo, M. & Kopeinig, S. (2008) Some practical guidance for the implementation of propensity score matching. *Journal of Economic Surveys*, 22, 31–72.
14. Linden, A. & Adams, J. L. (2008) Improving participant selection in disease management programs: insights gained from propensity score stratification. *Journal of Evaluation in Clinical Practice*, 14, 914–918.
15. Cochran, W. G. (1968) The effectiveness of adjustment by subclassification in removing bias in observational studies. *Biometrics*, 24, 205–213.
16. Rosenbaum, P. R. & Rubin, D. B. (1984) Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association*, 79, 516–524.
17. Rosenbaum, P. R. (1987) Model-based direct adjustment. *Journal of the American Statistical Association*, 82, 387–394.
18. Bang, H. & Robins, J. M. (2005) Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61, 962–973.
19. Wooldridge, J. M. (2010) Econometric Analysis of Cross Section and Panel Data, 2nd edn. Cambridge, MA: MIT Press.
20. Huang, I.-C., Frangakis, C., Dominici, F., Diette, G. B. & Wu, A. W. (2005) Application of a propensity score approach for risk adjustment in profiling multiple physician groups on asthma care. *Health Services Research*, 40, 253–278.
21. Hong, G. (2010) Marginal mean weighting through stratification: adjustment for selection bias in multilevel data. *Journal of Educational and Behavioral Statistics*, 35, 499–531.
22. Hong, G. (2012) Marginal mean weighting through stratification: a generalized method for evaluating multi-valued and multiple treatments with non-experimental data. *Psychological Methods*, 17, 44–60.
23. Linden, A. (2006) What will it take for disease management to demonstrate a return on investment? New perspectives on an old theme. *American Journal of Managed Care*, 12, 217–222.
24. Flury, B. K. & Reidwyl, H. (1986) Standard distance in univariate and multivariate analysis. *The American Statistician*, 40, 249–251.
25. Charlson, M. E., Pompei, P., Ales, K. L. & McKenzie, C. R. (1987) A new method of classifying prognostic comorbidity in longitudinal studies: development and validation. *Journal of Chronic Disease*, 40, 373–383.
26. Long, J. S. (1997) Regression Models for Categorical and Limited Dependent Variables. Thousand Oaks, CA: Sage.
27. Lu, B., Zanutto, E., Hornik, R. & Rosenbaum, P. R. (2001) Matching with doses in an observational study of a media campaign against drug abuse. *Journal of the American Statistical Association*, 96, 1245–1253.
28. Rubin, D. B. (1973) The use of matched sampling and regression adjustment to remove bias in observational studies. *Biometrics*, 29, 185–203.

29. Segal, J. B., Griswold, M., Achy-Brou, A., Herbert, R., Bass, E. B., Dy, S. M., Millman, A. E., Wu, A. W. & Frangakis, C. E. (2007) Using propensity scores subclassification to estimate effects of longitudinal treatments: an example using a new diabetes medication. *Medical Care*, 45, S149–S157.

30. Rosenbaum, P. (2002) Observational Studies, 2nd edn. New York, NY: Springer.

Supporting Information

Additional supporting information may be found in the online version of this article at the publisher's web site.