Using Machine Learning to Assess Covariate Balance in Matching Studies

Ariel Linden, DrPH[1,2], Paul R. Yarnold, PhD[3]

[1] President, Linden Consulting Group, LLC - Ann Arbor, MI alinden@lindenconsulting.org

[2] Research Scientist, Division of General Medicine, Medical School - University of Michigan, Ann Arbor, Michigan, USA

[3] President, Optimal Data Analysis, LLC

Corresponding Author Information:
Ariel Linden, DrPH
Linden Consulting Group, LLC
1301 North Bay Drive
Ann Arbor, MI USA 48103
Phone: (971) 409-3505
Email: alinden@lindenconsulting.org

Key Words: machine learning, data mining, matching studies, observed characteristics, covariate balance, selection bias

Running Header: machine learning for assessing covariate balance

Acknowledgement: We wish to thank Julia Adler-Milstein for reviewing the manuscript and providing many helpful comments.

This is the author manuscript accepted for publication and has undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1111/jep.12538

ABSTRACT

In order to assess the effectiveness of matching approaches in observational studies, investigators typically present summary statistics for each observed pre-intervention covariate, with the objective of showing that matching reduces the difference in means (or proportions) between groups to as close to zero as possible. In this paper, we introduce a new approach to distinguish between study groups based on their distributions of the covariates using a machine learning algorithm called optimal discriminant analysis (ODA). Assessing covariate balance using ODA as compared to the conventional method has several key advantages: the ability to ascertain how individuals self-select based on optimal (maximum-accuracy) cut-points on the covariates; the application to any variable metric and number of groups; its insensitivity to skewed data or outliers; and the use of accuracy measures that can be widely applied to all analyses. Moreover, ODA accepts analytic weights, thereby extending the assessment of covariate balance to any study design where weights are used for covariate adjustment. By comparing the two approaches using empirical data, we are able to demonstrate that using measures of classification accuracy as balance diagnostics produce highly consistent results to those obtained via the conventional approach (in our matched-pairs example ODA revealed a weak statistically significant relationship not detected by the conventional approach). Thus, investigators should consider ODA as a robust complement, or perhaps alternative, to the conventional approach for assessing covariate balance in matching studies.

1. INTRODUCTION

Although the randomized-controlled trial (RCT) is the gold standard for evaluating health interventions, the extent of its use is limited due to myriad practical, logistical and ethical reasons. Therefore, in circumstances when the RCT is not feasible, investigators typically choose from a wide variety of matching approaches in an attempt to emulate the randomization process using observational data [1]. The fundamental difference between the RCT and matching studies is that randomization is expected to produce study groups that are balanced (comparable) on both observed and unobserved pre-intervention characteristics, while matching studies can only strive to create study groups that are balanced on observed pre-intervention characteristics, and must assume that any unmeasured variables will not bias the results [2]. Consequently, demonstrating how well the study arms balance on their pre-intervention characteristics is an essential condition for making the case for the validity of treatment effects in matching studies.

The conventional approach to show comparability between study groups is to present a table of summary statistics for all observed pre-intervention covariates -- both before and after matching [3]. The objective is to simply demonstrate that matching reduces the difference in means (or proportions) between groups, to as close to zero as possible. However, this approach is sensitive to skewed data and outliers [4], limited to comparisons between two groups (unless multiple pairwise comparisons are made) [5], and perhaps most importantly, it does not identify a cut-point along the distribution of the covariate which may clarify how individuals self-select into one or the other study group.

In this paper, we describe a novel approach to assessing comparability between study groups in matching studies that overcomes the limitations of the difference-in-means diagnostic. This approach involves a machine-learning algorithm called optimal discriminant analysis (ODA) [6,7] that determines if (and to what degree) study groups can be distinguished based on the distributions of the covariates. The assumption is that individuals who elect to participate in observational studies generally differ in their characteristics from those who decline to participate and the algorithm should therefore be able to find characteristics that discriminate between groups prior to matching [8,9]. If matching is successful, reprocessing the algorithm on the matched groups should fail to identify characteristics that discriminate between the groups. This approach generates measures of classification accuracy (e.g., sensitivity, specificity, effect strength for sensitivity [9,10] as balance diagnostics, thereby providing additional information as to how well matching improved the comparability between study groups. The specific advantages of ODA as compared to the conventional approach in assessing covariate balance are; the application to any variable metric and number of groups, its insensitivity to skewed data or outliers, and the use of accuracy measures that can be widely applied to all analyses. ODA also has the distinct ability to ascertain how individuals self-select based on optimal (maximum-accuracy) cut-points on the covariates. Moreover, ODA accepts analytic weights, thereby extending the assessment of covariate balance to any study design where weights are used for covariate adjustment. Finally, ODA has the capability to use cross-validation in assessing the generalizability of the model, or to identify solutions that cross-generalize with maximum accuracy when applied across multiple samples.

To illustrate the ODA approach and compare it to the conventional method, the paper is organized as follows. In Section 2 we describe our methods including the data source, the matching methodology employed, a brief introduction to ODA, how each approach assesses covariate balance and how we compared the conventional approach and the ODA approach for assessing covariate balance. Section 3 reports the results of each approach and the comparison between them. Section 4 discusses the specific advantages of ODA in assessing covariate balance compared to the conventional approach; explains how machine-learning techniques like ODA might be incorporated into conventional methods for assessing covariate balance, selection and recruitment; and describes how machine-learning can be applied more broadly within the causal inferential framework.

2. METHODS

Data

Our empirical example uses data from a prior evaluation of a primary care-based medical home pilot program that invited patients to enroll if they had a chronic illness or were predicted to have high costs in the following year. The goal of the program was to lower healthcare costs for program participants by providing intensified primary care (see [11] for a more comprehensive description). The retrospectively collected data consist of observations for 374 program participants and 1,628 non-participants. Eleven pre-intervention characteristics were available; these included demographic variables (age and gender), health services utilization in year prior to enrollment (primary care visits, other outpatient visits, laboratory tests, radiology tests,

prescriptions filled, hospitalizations, emergency department visits, and home-health visits) and total medical costs (the amount paid for all those health services utilized in the prior year).

Propensity Score Matching

As described in Linden [11], a propensity score-based matching approach was employed to make the groups similar on observed baseline characteristics. The propensity score is defined as the probability of assignment to the treatment group given the observed characteristics [12] and we estimated the propensity score using the conventional approach of logistic regression to predict program participation status using the eleven pre-intervention covariates described above, all entered as main effects. It has been demonstrated that in large samples, when treatment and control groups have similar distributions of the propensity score, they generally have similar distributions of the underlying covariates used to create the propensity score. This means that observed baseline covariates can be considered independent of treatment assignment (as if they were randomized), and therefore will not bias the treatment effects [12]. To achieve this similar distribution of the propensity score in our study, an optimal matching algorithm [13] was employed to match pairs (one participant to one non-participating control) on the estimated propensity score, resulting in 276 matched pairs [11].

A Brief Introduction to Optimal Discriminant Analysis (ODA)

ODA is a machine learning algorithm that was introduced over 25 years ago [14] as an alternative means of analyzing data commonly encountered in research, such as studies with two or more study group levels and a variable of interest (e.g., a pre-intervention characteristic or outcome variable) that is measured on a continuous or interval-level scale, on an ordered scale

with relatively few levels, or on a qualitative scale with two or more categories [6,7,15]. In simple terms, ODA identifies the cutpoint (or category subset) of the variable of interest that yields maximum classification accuracy -- that is, the assignment rule that most accurately classifies observations into their actual study group. Maximum classification accuracy may be either overall percent accuracy in classification [PAC], or effect strength for sensitivity [ESS] (described in the next Section) depending on whether or not the investigator chooses to weight the data by prior odds [6,7]. For an ordered or continuous variable, the model has the form: if score $\leq$ (value) predict that the observation is from study group A; otherwise predict that the observation is from study group B. For a categorical variable, the model has the form: if score = (category list) predict the observation is from treatment group A; otherwise predict treatment group B. Statistical significance of the PAC and ESS statistics is evaluated using a permutation probability (no distributional assumptions are made) [7,9].

Conventional and ODA Approaches to Assessing Covariate Balance

After matching, the conventional approach to assessing whether matching successfully created covariate balance is to compare differences in means. The standardized difference [16] is perhaps the most widely used measure of balance and is simple to compute with data presented in a table of baseline characteristics:

$$SD = \frac{\left|\overline{X}_T - \overline{X}_C\right|}{\sqrt{\frac{(S_T)^2 + (S_C)^2}{2}}} \qquad (1)$$

where the numerator is the absolute difference in means between the treatment and control groups (denoted as *T* and *C*, respectively) and the denominator is a 50:50 pooled standard deviation [3]. Dichotomous covariates can also be tested for balance using this equation or using a formula specific to proportions [17]. While there is currently no universally-recognized cut-off point as to what is considered the upper limit of balance, Normand et al. [18] suggest that a standardized difference of less than 0.10 is indicative of good balance.

Assessing covariate balance using ODA involves three measures of accuracy. *Sensitivity* (true positive rate) is the proportion of actual participants that are correctly predicted by the ODA model as being participants. *Specificity* (true negative rate) is the proportion of actual non-participants that are correctly predicted by the ODA model as being non-participants. Finally, a measure of accuracy that combines these two metrics is the effect strength for sensitivity (ESS), introduced by Yarnold and Soltysik [6]. ESS is a chance-corrected (0 = the level of accuracy expected by chance) and maximum-corrected (100 = perfect prediction) index of predictive accuracy. The formula for computing ESS for binary case classification is:

ESS = [(Mean Percent Accuracy in Classification –50)]/ 50 x 100% (2),

where

Mean Percent Accuracy in Classification = (sensitivity + specificity)/2 x 100 (3).

Yarnold and Soltysik [6] consider ESS values less than 25% to indicate a relatively weak, 25% to 50% to indicate a moderate, 50% to 75% to indicate a relatively strong, and 75% or greater to indicate a strong effect. Using ESS, an investigator may directly compare the

performance among the various covariates -- pre- and post-matching, regardless of structural features of the analyses, such as sample size and the measurement metric.

Comparing Conventional and ODA Approaches to Assessing Covariate Balance

Given that the two approaches rely on different estimators of covariate balance, we compare them by observing whether they agree that covariate balance has been achieved, and if not, whether they identify the same covariate(s) as being imbalanced. In determining whether there is agreement on whether balance has been achieved, we compare *P* values for each of the individual covariates that are produced by each of the approaches. In the conventional approach, *P* values for continuous variables were estimated using a two-tailed t-test for independent samples (t-tests for matched samples were used in the matched pairs analysis), and the Chi-squared test was used for dichotomous variables. In all ODA models, *P* values were calculated using permutation tests, which are estimates derived from 25,000 Monte Carlo simulations [7,19]. To the extent that *P* values are of similar magnitude, they suggest consistency between the two approaches. To the extent that they are different, it may suggest that one approach is more sensitive to detecting differences.

3. RESULTS

Table 1 presents the observed pre-intervention characteristics of the participants and non-participants [11]. Continuous variables are summarized by the mean and standard deviation, and categorical variables are presented as number and percent. For balance measures, we report the absolute standardized difference -- for which perfect balance is zero, and the conventional *P* value, where variables with values ≤ 0.05 may be considered imbalanced. It is clear that the

participant group differed markedly from the non-participant group on every covariate. On average, participants were older, were less likely to be female, and overall had higher utilization and costs than non-participants. All standardized differences were substantially greater than zero, and all *P* values were ≤ 0.05.

Table 2 displays the baseline characteristics of the participants and their propensity score matched controls as would be presented in the conventional approach. It is evident from reviewing the absolute standardized differences that the matching procedure was successful in reducing imbalances of all observed baseline covariates to under 0.10, and all *P* values were much greater than > 0.05. Thus, by the conventional method, the two groups would be considered balanced on all observed pre-intervention characteristics.

Table 3 presents the observed pre-intervention characteristics of the participants and non-participants analyzed using ODA. Summary values represent the cutoff point on the covariate, sensitivity is presented for participants, and specificity is presented for non-participants. For balance measures, we report the ESS (for which higher percentage values represent better classification accuracy and ability to discriminate between groups), and the *P* value derived using permutation tests, where variables with *P* values ≤ 0.05 may be considered imbalanced.

To help interpret the ODA results, we use the covariate *age* as an example of a continuous variable and the covariate *female* as an example of a categorical variable. The ODA model predicted that an individual was a participant in the study if their age was greater than 49.5, and a non-participant if their age was less than or equal to 49.5. The ODA model correctly classified 79.95% of participants and 63.70% of non-participants according to their age (Table

3). Classification performance was moderate (ESS = 43.64%) and statistically significant (permuted P <0.001). Thus, the results for age using ODA are consistent with those using the conventional approach (Table 1), that is, higher age is predictive of participation in the pilot while lower age is predictive of non-participation. For the covariate *female*, the ODA model predicted that an individual was a participant in the study if they were a female and a non-participant if they were a male. The ODA model was only able to correctly classify 56.42% of participants and 50.43% of non-participants based on their gender. The classification performance was weak (ESS = 6.85%) in practical terms, yet statistically significant (permuted P <0.001). Thus, although a statistically greater proportion of females versus males were in the treatment group, the practical strength of this difference is only marginally greater than chance. The results for all other covariates are interpreted analogously.

Table 4 displays the baseline characteristics of the participants and their propensity score-matched controls, analyzed using ODA. As in Table 3, summary values represent the cutoff point on the covariate, sensitivity is presented for participants, and specificity is presented for matched controls. As a whole, ODA had difficulty in accurately predicting treatment assignment in both study groups for all of the covariates under study. In some cases, the cutoff point leads to greater sensitivity at the expense of lower specificity (e.g. primary care visits), while in other cases, the cutoff point leads to greater specificity at the expense of lower sensitivity (e.g. hospitalizations). As a consequence, the combined classification performance (as measured via ESS) is consistently weak across all covariates, and is supported by non-statistically significant P values > 0.05.

In comparing the two approaches for the matched pairs, we see that both produced consistent results concerning the weak differences between groups: neither approach identified any covariates that were strongly imbalanced. However, when comparing the results by reviewing *P* values, we see that for prescriptions filled, the conventional approach estimated a *P* value of 0.516, versus 0.027 by ODA. This reflects the fact that in the conventional approach involving the comparison of means, the validity of the *P* values depends on the underlying distributional assumptions being met. For the ODA approach involving maximizing predictive accuracy, exact *P* values do not require any distributional assumptions. Thus, an advantage of ODA is that the validity of the *P* values is guaranteed, and does not need to be evaluated. Furthermore, the ESS and corresponding *P* value for ODA is invariant over any monotonic transformation of the variable [7,20]. Generalizing this situation, it suggests that ODA will generally do a better job of detecting differences for covariates that do not well adhere to their underlying distributional assumptions. Of note, the ESS for each covariate in Tables 3 and 4 appears to agree reasonably well with the corresponding SD values in Tables 1 and 2. For example the highest SD values are for age, primary care visits, and prescriptions filled (Table 1). The ESS values for these covariates (Table 3) are similarly high. Likewise, home-health visits have both a low SD and a low ESS (indicating good covariate balance in the conventional method and analogously, poor discriminatory ability using ODA).

To summarize, we see that in this example, the determination of covariate balance from ODA is consistent with findings from the conventional approach. When participants are compared to non-participants, both methods are able to distinguish between study groups – an

expected finding, given that individuals self-select into observational studies and are likely to have different characteristics than those who elect not to participate. Similarly, when the intervention group is compared to the matched control group, neither method is able to distinguish between study groups on their pre-intervention characteristics (if the study corrects for multiple comparisons, then the results of the statistical significance findings of both methods are isomorphic [7]). This result indicates that matching generated comparable study groups based on the observed pre-intervention characteristics.

4. DISCUSSION

Given that ODA provided similar results to those using the conventional comparison of means approach in our example, one may question the need for an additional method for assessing covariate balance in observational studies. Of course this is not always the case -- conventional and ODA analyses have obtained strongly divergent findings in a wide variety of real-world data [6,7]. Nevertheless, even in applications for which conceptually parallel conclusions regarding strength and statistical reliability are obtained, ODA offers key advantages specifically for assessing covariate balance that cannot be realized using the conventional approach alone. For example, the ODA algorithm, with its associated measure of classification performance (ESS) and non-parametric permutation tests, can be universally applied to any variable type and number of study groups, and is not affected by skewed data or outliers -- a concern that may arise in the context of meeting assumptions underlying the validity of the estimated *P* value

using the conventional approach alone (for example, as is evident in the current data by the large standard deviations for most covariates in Table 1).

Beyond those advantages, ODA can also help explain how individuals self-select in observational studies, by identifying group membership based on the cut-point on the covariate (i.e., category subset). So for example, while we can only say that *on average*, participants spent $5,189 more on healthcare than non-participants in the year prior to the study (Table 1), ODA provides more specific information -- i.e. individuals with medical costs > $2,773 were much more likely (with a relatively strong effect strength) to participate in the study than individuals with costs ≤ $2,773 (Table 3). This level of precision (cut-point) and classification detail (model sensitivity and specificity) provides health researchers with a better understanding of the selection process in observational studies than is possible by a simple comparison of means. This may be useful, for example, if investigators would like to create a tailored recruitment plan targeting individuals who are most likely to benefit from the intervention [7,9].

Moreover, because ODA allows the use of analytic weights, the algorithm can be extended the assessment of covariate balance to any study design where weights are used for covariate adjustment (see for example [5,6,7,21,22,23,24,25]).

Finally, ODA can be implemented using cross-validation to assess the generalizability of the model, such as *k*-fold cross-validation, bootstrapping, and leave-one-out jackknife cross-validation [6,7,26,27]. This typically entails first estimating a model using the entire sample (training set) and calculating the accuracy measures, followed by the same model being tested on one or more hold-out (test) samples and then recalculating the accuracy measures. If the

accuracy measures remain consistent with those of the original model using the entire sample, then the model is considered generalizable. This may important, for example, if the goal of the analysis is to assist health researchers identify new candidates for participation in an ongoing intervention, or initiate the intervention in other settings [9,28]. Cross-validation is less important if the goal is only to estimate treatment effects of the intervention.

In applications involving two or more independent samples, ODA may also be used to identify a model that explicitly maximizes classification accuracy across the multiple samples using the "Gen" (for generalizability) algorithm [6,7,29]. Using this methodology, ODA identifies that model which, when simultaneously and independently applied to each of the samples, maximizes the minimum ESS achieved across the samples. If the resulting level of classification performance meets or exceeds the researcher's *a priori* specification for acceptable performance, then the single model may be used to classify observations in all of the samples. If one or more samples yield findings that fail to meet the *a priori* specification, then it is concluded that one model cannot be used with every sample. Samples for which inadequate performance was obtained can be eliminated, and the analysis reprocessed. In this manner the multiple samples can be separated using a minimum subset of models that achieve satisfactory performance in the context of the hypothesis being tested.

A major limitation of the conventional and ODA approaches is that covariate balance is assessed in only a single dimension, possibly leaving imbalances at other points in the distribution [4]. In the conventional approach, this may be addressed (to some degree) by assessing balance at other moments in the distribution and interactions. In the maximum-

accuracy statistical paradigm, this can be remedied by using optimal classification tree analysis (CTA) -- that involves recursive partitioning using chained ODA analyses -- to identify a nonlinear model for discriminating the groups on the basis of the covariates [7,9,30,31,32]. Applying this methodology presently was unproductive, because no additional classification accuracy was possible beyond the variable identified using ODA (number of prescriptions). This approach demonstrates that there are no additional subsets of observations on which additional covariates differ. However, if CTA identified differences on two or more covariates, then the resulting sample strata (groups) that differed with respect to specific combinations of covariates would be specifically identified vis-à-vis cut-points (or category lists), and information regarding the strength (ESS), reliability (*P* value), and cross-generalizability (validity analysis) of the differences would be reported.

While this paper has focused solely on using machine learning algorithms to assess covariate balance, more broadly, there are several additional aspects in the evaluation of observational studies where machine learning techniques can be applied. For example, Linden and Yarnold [9] use CTA to characterize the nature of individuals who choose to participate in observational studies, while Athey & Imbens [33] modify the conventional classification and regression trees (CART) approach to estimate heterogeneous causal effects in such studies. One can also envision the use of such classification algorithms to identify potential instrumental variables that may provide an unbiased estimate of the causal effect of intervention on the outcome (IV). An IV is a variable (*Z*) that is correlated with the intervention (*X*), but not associated with unobserved confounders of the outcome (*Y*) [34]. Potential IVs may be identified

by first generating a CTA model predicting participation (as in [9]) and then generating a second model predicting the outcome -- allowing the same set of covariates in both models. Covariates that appear in the first (selection) model, but not in the second (outcome) model, may be suggestive of potential IVs, which can then be used within the IV framework. In general, the application of machine-learning techniques to improve causal inference in observational studies is open to much further exploration. And in particular, emphasis should be placed on determining the most appropriate algorithm for a given problem -- or a generalization to all algorithms, extension to outcomes with censored data [35], and the development of specific sensitivity analyses for these applications [36].

In summary, ODA can serve as a complement, or as an alternative, to the conventional approach for testing covariate balance, providing additional dimensions and robustness to the analysis that may help with issues related to selection and recruitment. More broadly, health researchers should consider the use of machine learning algorithms to improve causal inference in observational studies by identifying patterns in the data that distinguish study participants from non-participants, and controlling for potentially complex relationships among individual characteristics that may bias the outcome analysis.

REFERENCES

1. Stuart, E.A. (2010) Matching methods for causal inference: a review and a look forward. *Statistical Science,* 25, 1–21.
2. Rubin, D.B. (1973) Matching to remove bias in observational studies. *Biometrics,* 29, 159–184.
3. 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.
4. Linden A**.** (2015) Graphical displays for assessing covariate balance in matching studies. *Journal of Evaluation in Clinical Practice*, 21, 242-247.
5. Linden, A., Uysal, S.D., Ryan, A., & Adams, J.L. (2016) Estimating causal effects for multivalued treatments: A comparison of approaches. *Statistics in Medicine,* 35, 534-552.
6. Yarnold, P.R., & Soltysik, R.C. (2005) *Optimal data analysis: A Guidebook with Software for Windows* Washington, DC: APA Books.
7. Yarnold, P.R., & Soltysik, R.C. (2016) *Maximizing Predictive Accuracy*. Chicago, IL: ODA Books. DOI: 10.13140/RG.2.1.1368.3286
8. Linden, A., Adams, J., & Roberts, N. (2003) An assessment of the total population approach for evaluating disease management program effectiveness. *Disease Management,* 6, 93-102.
9. Linden, A., & Yarnold, P.R. (*In Print*) Using data mining techniques to characterize participation in observational studies. *Journal of Evaluation in Clinical Practice*.

10. Linden, A. (2006) Measuring diagnostic and predictive accuracy in disease management: an introduction to receiver operating characteristic (ROC) analysis. *Journal of Evaluation in Clinical Practice,* 12, 132-139.
11. Linden, A. (2011) Identifying spin in health management evaluations. *Journal of Evaluation in Clinical Practice,* 17, 1223-1230.
12. Rosenbaum, P.R., & Rubin, D.B. (1983) The central role of propensity score in observational studies for causal effects. *Biometrika,* 70, 41–55.
13. Rosenbaum, P.R. (1989) Optimal matching for observational studies. *Journal of the American Statistical Association,* 84, 1024-1302.
14. Yarnold, P.R., & Soltysik, R.C. (1991). Theoretical distributions of optima for univariate discrimination of random data. *Decision Sciences, 22,* 739-752.
15. Carmony, L., Yarnold, P.R., & Naeymi-Rad, F. (1997). One-tailed Type I error rates for balanced two-category UniODA with a random ordered attribute. *Annals of Operations Research, 74,* 223-238.
16. Flury, B.K. & Reidwyl, H. (1986) Standard distance in univariate and multivariate analysis. *The American Statistician,* 40, 249–251.
17. Austin, P.C. (2009**)** Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. *Statistics in Medicine,* 28, 3083-3107.
18. Normand, S.L.T., Landrum, M.B., Guadagnoli, E., Ayanian, J.Z., Ryan, T.J., Cleary, P.D., & McNeil, B.J. (2001) Validating recommendations for coronary angiography following an

acute myocardial infarction in the elderly: a matched analysis using propensity scores. *Journal of Clinical Epidemiology*, 54, 387–398.

19. Yarnold, P.R., & Soltysik, R.C. (2010). Precision and convergence of Monte Carlo Estimation of two-category UniODA two-tailed *p*. *Optimal Data Analysis*, *1*, 43-45.
20. Yarnold, P.R., Soltysik, R.C., & Martin, G.J. (1994). Heart rate variability and susceptibility for sudden cardiac death: An example of multivariable optimal discriminant analysis. *Statistics in Medicine*, *13*, 1015-1021.
21. Linden, A., & Adams, J.L. (2010a) Using propensity score-based weighting in the evaluation of health management programme effectiveness. *Journal of Evaluation in Clinical Practice*, 16, 175-179.
22. Linden, A., & Adams, J.L. (2010b) Evaluating health management programmes over time. Application of propensity score-based weighting to longitudinal data. *Journal of Evaluation in Clinical Practice*, 16, 180-185.
23. Linden, A., & Adams, J.L. (2011) Applying a propensity-score based weighting model to interrupted time series data: improving causal inference in program evaluation. *Journal of Evaluation in Clinical Practice*, 17, 1231-1238.
24. Linden, A., & Adams, J.L. (2012) Combining the regression-discontinuity design and propensity-score based weighting to improve causal inference in program evaluation. *Journal of Evaluation in Clinical Practice*, 18, 317-325.

25. Linden, A. (2014) Combining propensity score-based stratification and weighting to improve causal inference in the evaluation of health care interventions. *Journal of Evaluation in Clinical Practice*, 20, 1065-1071.

26. Witten, I.H., Frank, E., & Hall, M.A. (2011) *Data Mining: Practical Machine Learning Tools and Techniques, 3rd edition*. San Francisco: Morgan Kaufmann.

27. Linden, A., Adams, J., & Roberts, N. (2005) Evaluating disease management program effectiveness: An introduction to the bootstrap technique. *Disease Management and Health Outcomes*, 13, 159-167.

28. Linden, A., Adams, J., & Roberts, N. (2004) The generalizability of disease management program results: getting from here to there. *Managed Care Interface*, 17, 38-45.

29. Yarnold, P.R. (2010). GenUniODA *vs*. log-linear model: Modeling discrimination in organizations. *Optimal Data Analysis, 1*, 59-61.

30. Yarnold, P.R. (1996). Discriminating geriatric and non-geriatric patients using functional status information: An example of classification tree analysis via UniODA. *Educational and Psychological Measurement, 56*, 656-667.

31. Yarnold, P.R., & Bryant, F.B. (2015a). Obtaining a hierarchically optimal CTA model via UniODA software. *Optimal Data Analysis, 4*, 36-53.

32. Yarnold, P.R., & Bryant, F.B. (2015b). Obtaining an enumerated CTA model via automated CTA software. *Optimal Data Analysis, 4*, 54-60.

33. Athey, S., & Imbens, G. (2015) Machine Learning Methods for Estimating Heterogeneous Causal Effects. *Working Paper*. Downloadable from http://arxiv.org/abs/1504.01132 [Accessed on 30 November 2015].
34. Linden, A., & Adams, J. (2006) Evaluating disease management program effectiveness: an introduction to instrumental variables. *Journal of Evaluation in Clinical Practice,* 12, 148-154.
35. Linden, A., Adams, J., & Roberts, N. (2004) Evaluating disease management program effectiveness: An introduction to survival analysis. *Disease Management*, 7, 180-190.
36. Linden, A., Adams, J., & Roberts, N. (2006) Strengthening the case for disease management effectiveness: unhiding the hidden bias. *Journal of Evaluation in Clinical Practice,* 12, 140-147.

Table 1: Baseline (12 months) characteristics of program participants and non-participants (from Linden [2011]). Continuous variables are reported as mean (standard deviation) and dichotomous variables are reported as *N* (percent).

	Participants (*N*=374)	Non-Participants (*N*=1628)	Standardized difference	*P*-value
Demographic characteristics				
Age	54.9 (6.71)	43.4 (11.99)	1.704	<0.001
Female	211 (56.4%)	807 (49.6%)	0.138	0.017
Utilization and Cost				
Primary care visits	11.3 (7.30)	4.6 (4.35)	0.914	<0.001
Other outpatient visits	18.0 (16.65)	7.2 (10.61)	0.647	<0.001
Laboratory tests	6.1 (5.27)	2.4 (3.31)	0.705	<0.001
Radiology tests	3.2 (4.46)	1.3 (2.48)	0.424	<0.001
Prescriptions filled	40.6 (29.96)	11.9 (17.14)	0.956	<0.001
Hospitalizations	0.2 (0.52)	0.1 (0.29)	0.326	<0.001
Emergency department visits	0.4 (1.03)	0.2 (0.50)	0.226	<0.001
Home-health visits	0.1 (0.88)	0.0 (0.38)	0.083	0.012
Total costs	8236 (9830)	3047 (5817)	0.528	<0.001

Table 2: Comparison of baseline characteristics of program participants and their 1:1 propensity score matched controls. Continuous variables are reported as mean (standard deviation) and dichotomous variables are reported as *N* (percent)

	Participants (*N*=276)	Matched Controls (*N*=276)	Standardized difference	*P*-value
Demographic characteristics				
Age	54.6 (6.5)	54.0 (6.9)	0.082	0.316
Female	152 (55.1%)	150 (54.3%)	0.015	0.864
Utilization and Cost				
Primary care visits	9.5 (6.5)	9.7 (6.2)	0.022	0.803
Other outpatient visits	15.2 (16.2)	15.6 (14.1)	0.029	0.751
Laboratory tests	4.8 (5.8)	5.2 (4.5)	0.086	0.380
Radiology tests	2.8 (4.4)	2.8 (4.1)	0.009	0.920
Prescriptions filled	32.6 (27.8)	34.1 (25.3)	0.058	0.516
Hospitalizations	0.2 (0.4)	0.2 (0.4)	0.026	0.768
Emergency department visits	0.3 (0.8)	0.3 (0.9)	0.027	0.729
Home-health visits	0.1 (0.9)	0.1 (1.0)	0.011	0.894
Total costs	6318 (7827)	6748 (7648)	0.056	0.513

Table 3: Baseline (12 months) characteristics of program participants and non-participants (from Linden [2011]). Values represent cut-points on the covariate, and values in parentheses represent sensitivity (for participants) and specificity (for non-participants).

	Participants (*N*=374)	Non-Participants (*N*=1628)	Effect Strength Sensitivity	*P*-value
Demographic characteristics				
Age	> 49.5 (79.95)	<= 49.5 (63.70)	43.64%	<0.001
Female	= 1 (56.42)	= 0 (50.43)	6.85%	0.020
Utilization and Cost				
Primary care visits	> 7.5 (67.38)	<= 7.5 (82.68)	50.06%	<0.001
Other outpatient visits	> 6.5 (75.13)	<= 6.5 (68.86)	43.99%	<0.001
Laboratory tests	> 2.5 (78.07)	<= 2.5 (67.38)	45.46%	<0.001
Radiology tests	> 1.5 (64.44)	<= 1.5 (69.96)	34.40%	<0.001
Prescriptions filled	> 16.5 (80.75)	<= 16.5 (77.09)	57.84%	<0.001
Hospitalizations	> 0.5 (19.25)	<= 0.5 (94.16)	13.42%	<0.001
Emergency department visits	> 0.5 (22.99)	<= 0.5 (88.45)	11.45%	<0.001
Home-health visits	> 2.5 (1.60)	<= 2.5 (99.82)	1.42%	0.002
Total costs	> 2773 (85.03)	<= 2773 (71.74)	56.77%	<0.001

Table 4: Comparison of baseline characteristics of program participants and their 1:1 propensity score matched controls. Values represent cut-points on the covariate, and values in parentheses represent sensitivity (for participants) and specificity (for non-participants).

	Participants (*N*=276)	Matched Controls (*N*=276)	Effect Strength Sensitivity	*P*-value
Demographic characteristics				
Age	<= 52.5 (37.68)	>52.5 (70.29)	7.97%	0.215
Female	= 0 (45.65)	= 1 (55.07)	0.72%	0.932
Utilization and Cost				
Primary care visits	> 2.5 (96.38)	<= 2.5 (9.06)	5.43%	0.590
Other outpatient visits	> 5.5 (74.64)	<= 5.5 (35.51)	10.14%	0.072
Laboratory tests	> 3.5 (59.06)	<= 3.5 (49.64)	8.70%	0.115
Radiology tests	> 0.5 (80.80)	<= 0.5 (24.28)	5.07%	0.519
Prescriptions filled	> 16.5 (76.09)	<= 16.5 (35.87)	11.96%	0.027
Hospitalizations	> 0.5 (14.86)	<= 0.5 (87.32)	2.17%	0.546
Emergency department visits	> 0.5 (21.38)	<= 0.5 (84.06)	5.43%	0.123
Home-health visits	> 2.5 (1.45)	<= 2.5 (98.91)	0.36%	0.847
Total costs	> 4629 (49.64)	<= 4629 (61.23)	10.87%	0.079