Combining machine learning and matching techniques to improve causal inference in program evaluation

Ariel Linden, DrPH[1,2], Paul R. Yarnold, PhD[3,4]

[1] President, Linden Consulting Group, LLC, Ann Arbor, Michigan, USA

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

[3] Principal Scientist, Optimal Data Analysis, LLC

[4] Statistician, Southern Network on Adverse Reactions (SONAR), College of Pharmacy, University of South Carolina, Columbia, South Carolina, USA

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, matching, balance, propensity score, causal inference

Running Header: machine learning and matching for causal inference

Acknowledgement: We wish to thank Julia Adler-Milstein for her review and feedback on the manuscript. Dr. Yarnold's work was partly funded by a grant from the National Cancer Institute (1R01CA165609-01A1).

ABSTRACT

Rationale, aims and objectives: Program evaluations often utilize various matching approaches to emulate the randomization process for group assignment in experimental studies. Typically, the matching strategy is implemented and then covariate balance is assessed before estimating treatment effects. This paper introduces a novel analytic framework utilizing a machine learning algorithm called optimal discriminant analysis (ODA) for assessing covariate balance and estimating treatment effects, once the matching strategy has been implemented. This framework holds several key advantages over the conventional approach: application to any variable metric and number of groups; insensitivity to skewed data or outliers; and use of accuracy measures applicable to all prognostic analyses. Moreover, ODA accepts analytic weights, thereby extending the methodology to any study design where weights are used for covariate adjustment or more precise (differential) outcome measurement.

Method: One-to-one matching on the propensity score was used as the matching strategy. Covariate balance was assessed using standardized difference in means (conventional approach) and measures of classification accuracy (ODA). Treatment effects were estimated using ordinary least squares regression and ODA.

Results: Using empirical data, ODA produced results highly consistent with those obtained via the conventional methodology for assessing covariate balance and estimating treatment effects.

Conclusions: When ODA is combined with matching techniques within a treatment effects framework, the results are consistent with conventional approaches. However, given that it

provides additional dimensions and robustness to the analysis versus what can currently be achieved using conventional approaches, ODA offers an appealing alternative.

INTRODUCTION

In experimental studies with two arms (treatment and control), outcomes may be analyzed by simply regressing the outcome on a treatment group indicator variable in order to estimate treatment effects. This minimal design is sufficient to provide unbiased treatment effect estimates when subjects are randomized. However, when analyzing non-randomized observational data, investigators estimate treatment effects by applying causal-inferential methods to control for threats to validity [1]. When evaluating health management programs, selection bias is a particularly prominent threat to validity because individuals with high levels of health care utilization or costs are specifically targeted for enrollment. Given their high outlier status at baseline, these individuals' outcomes are naturally likely to regress to the mean on their follow-up measurement, giving the false impression of a treatment effect [2,3].

In observational studies, investigators typically choose from a wide variety of matching approaches in an attempt to emulate the randomization process for group assignment using observational data [4]. However, unlike most experimental studies in which study groups are inherently comparable on both observed and unobserved pre-intervention characteristics, matching studies can only endeavor to generate study groups that are comparable on observed characteristics and must assume that any unmeasured variables will not bias the results [5]. Thus, in evaluating a health management program using a matching approach, the investigator would ensure that study groups were comparable on pre-intervention levels of health care utilization and cost, but must assume, for example, that unmeasured motivation to change health behaviors will not confound the outcomes [6,7]. Accordingly, an essential condition for assuming the validity of treatment effects in matching studies is that the study groups are comparable on their observed pre-intervention characteristics [8,9].

Recently, Linden & Yarnold [10] introduced a novel machine-learning approach for assessing comparability between study groups in matching studies. This methodology employs an algorithm called optimal discriminant analysis (ODA) to determine if, and to what degree, study groups can be distinguished based on the distributions of the covariates [11,12]. Matching is considered successful if the ODA algorithm fails to identify characteristics that discriminate between the groups.

In this paper we extend this machine-learning approach to program evaluations that use matching as the basis for ensuring comparability between study groups. By framing the treatment-outcome relationship as a classification problem (i.e., how accurately does the outcome variable classify individuals as being in the treatment or control group) ODA offers several benefits over the conventional statistical methods typically employed in matching studies. These include the ability to handle an outcome variable measured using any metric (from categorical to continuous) and multiple treatment groups, insensitivity to skewed data or outliers, and the use of accuracy measures that can be widely applied to all classification analyses. ODA also offers the unique ability to ascertain if individuals are likely to be responding to the treatment as assigned (or self-selected) based on optimized (maximum-accuracy) cut-points on the outcome variable. Moreover, ODA accepts analytic weights, thereby allowing the evaluation of observational studies using any matching algorithm that produces weights for covariate adjustment [10]. Finally, ODA provides the capability to use cross-validation in assessing the generalizability of the model to other individuals outside of the original study sample, or to identify solutions that cross-generalize with maximum accuracy when applied across multiple samples [12].

To illustrate the ODA-matching framework, and compare it to the conventional method, the paper is organized as follows. In the Methods section we provide a brief introduction to ODA, and describe the data source and analytic framework employed in the current study. The Results section reports and compares the results of the conventional approach and ODA-matching framework. The Discussion section describes the specific advantages of ODA-matching framework for assessing covariate balance and evaluating treatment effects compared with the conventional approach, and discusses how machine-learning can be applied more broadly within the causal inferential framework.

METHODS

A brief introduction to optimal discriminant analysis

ODA is a machine learning algorithm that was introduced over 25 years ago to offer an alternative analytic approach to conventional statistical methods commonly used in research [13]. Its appeal lies in its simplicity, flexibility, and accuracy as compared to conventional methods [12,14,15].

To briefly describe how an ODA model is obtained, assume we have a continuous outcome (attribute) and a binary treatment (class) variable. First, we order the outcome variable from low to high. Next, we find all the points along the continuum of the outcome in which the next value belongs to an individual from the alternate class than that of the previous value (e.g. the next value belongs to a treated subject whereas the previous value belongs to a control). The *cutpoint* thus represents the mean value of the outcome at this point: cutpoint = (previous value + current value) / 2. *Directionality* defines how cutpoints are used to classify individual observations. The two directions are "less than" (controls have lower values on the outcome than treated subjects), and "greater than" (controls have higher values on the outcome than treated subjects). For an exploratory "two-tailed" hypothesis (controls and treated subjects have different values on the outcome), both directions are evaluated by the ODA algorithm. For a confirmatory "one-tailed" hypothesis (controls have lower values), only the appropriate direction (less than) is evaluated. For each cutpoint along the continuum of the outcome, ODA assesses how well the model—that is, the combination of cutpoint and direction—correctly predicts (in the current example) that controls have values of the outcome less than or equal to the cutpoint, and treated subjects have values of the outcome greater than the cutpoint [12,13].

ODA relies on three measures of accuracy to identify the optimal (maximum-accuracy) model – that is, the exact combination of cutpoint and direction that produces the most accurate predictions possible for the sample. *Sensitivity* or true positive rate is the proportion of actual treated subjects that are correctly predicted by the ODA model -- that is, those who have a value on the outcome that lies above the cutpoint. *Specificity* or true negative rate is the proportion of actual control subjects that are correctly predicted by the ODA model – that is, those who have a value on the outcome that lies at or below the cutpoint [16]. The third measure of accuracy combines these two metrics and is called the effect strength for sensitivity or ESS [11,12]. ESS is a chance-corrected (0 = the level of accuracy expected by chance) and a maximum-corrected (100 = perfect prediction) index of predictive accuracy. The formula for computing ESS for a binary (two-category) case classification result is:

ESS = [(Mean Percent Accuracy in Classification –50)]/ 50 x 100% (1),

where

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

The ODA algorithm iterates through each successive cutpoint and calculates ESS. The maximally-accurate model is that which has the cutpoint and direction with the highest associated value of ESS. Based on simulation research, ESS values <25% conventionally indicate a relatively weak, <50% indicate a moderate, 50-75% indicate a relatively strong, and ≥75% indicate a strong effect [11,12].

ODA also computes *P*-values to assess the statistical reliability (or “significance”) of the maximally-accurate ODA model. *P*-values are estimated using Monte Carlo permutation experiments. For example, in models with a binary treatment, this involves repeatedly shuffling subjects’ treatment assignment at random, holding their outcome value fixed at its true value. In each permuted dataset the ESS is recorded, and the permutation *P*-value represents the proportion of all permuted datasets in which the ESS is higher than the ESS of the maximally-accurate ODA model [11,12,13].

Finally, ODA can be implemented using cross-validation to assess the generalizability of the model, using methods such as *k*-fold cross-validation, bootstrapping, and leave-one-out jackknife cross-validation [18,17,12]. This typically entails first estimating a model using a training sample and calculating the accuracy measures, followed by applying the same model to 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 be 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. Cross-validation is less important if the goal is only to estimate treatment effects of the existing intervention [10,19,20].

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 [21] 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).

Analytic approach

Matching studies involve a sequential process in which the chosen matching strategy is implemented, covariate balance is assessed (i.e., comparability between study groups on pre-intervention characteristics), and then treatment effects are estimated [4,8,9]. This approach serves to emulate an experimental study by decoupling the design phase (matching on pre-intervention characteristics without consideration of the outcome as would be the case in a prospective study) from the analysis [22].

While any matching algorithm could be employed within our proposed machine-learning framework, for the purpose of this empirical example, a one-to-one, propensity score based matching approach was used, as implemented in Linden [21]. The propensity score is defined as the probability of assignment to the treatment group given the observed characteristics [23]. To be consistent with prior research we estimated the propensity score via the conventional approach of using 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 [23]. To achieve this similar distribution of the propensity score, an optimal matching algorithm [24] was used to match pairs (one participant to one non-participating control) on the estimated propensity score, resulting in 276 matched pairs [21].

Next, the effectiveness of the matching approach in reducing bias was examined by assessing covariate balance, comparing the conventional method to ODA: the conventional method compares the standardized difference in means [25], and ODA utilizes the aforementioned measures of accuracy – sensitivity, specificity, and ESS [11,12,16]. The expectation is that a well-matched cohort will have standardized differences close to zero, and poor (i.e., low) measures of accuracy [10].

The outcome analysis was performed by regressing health care costs in the program year on the treatment variable using ordinary least squares regression (OLS) with robust standard errors [26,27], and an exploratory ODA model was obtained in which health care costs was used as the attribute and treatment as the class variable, without specifying *a priori* directionality (i.e., hypothesizing a positive or negative difference over time). Exact *P* values were estimated using 25,000 Monte Carlo experiments [12].

Both analyses were performed on the unmatched population (naïve estimate) and on the matched sample (adjusted), in order to assess the degree to which matching reduced confounding and altered the treatment effect estimates.

Stata 14.1 (StataCorp., College Station, TX, USA) was used to conduct all conventional statistical analyses (i.e. covariate balance and outcome analyses using OLS regression), and ODA analyses were performed using UniODA Software [11].

RESULTS

Tables of covariate balance, before and after matching, are replicated from Linden [21] and Linden & Yarnold [10] and are presented in Tables 1-4. Both conventional and ODA methods found that matching generated comparable study groups based on the observed pre-intervention characteristics. In the former, standardized differences close to zero and *P*-values >

0.05 (Table 2). In the latter, we found consistently weak ESS values with permuted P values > 0.05 except for prescriptions filled ($P \le 0.027$, Table 4).

Table 5 presents program year costs (outcome) for participants and non-participants, both unadjusted and after one-to-one matching, using conventional OLS regression. As shown, the naïve treatment effect estimate (not controlling for confounding) is $4,038 (95% CI: 2,922, 5,154). In other words, while the treatment group is estimated to have statistically significantly higher ($P < 0.0001$) health care costs in the program year than the unmatched pool of non-participants, the regression model fails to explain the vast majority of the variance in health care costs between the two groups (97.59% remains unexplained). However, after controlling for confounding via matching, the subset of treated subjects was estimated to have program year costs that was, on average, $1,501 lower than that of the control group – but that was statistically indistinguishable from the control group ($P < 0.193$; 95% CI: -3,762, 760).

Table 6 presents program year costs (outcome) for participants and non-participants, both unadjusted and after one-to-one matching, using ODA as the analytic tool. Summary values represent the cutoff point on the outcome, sensitivity is presented for participants, and specificity is presented for non-participants. The ESS is reported as a measure of "clinical" importance (for which higher percentage values represent better classification accuracy and ability to discriminate between groups), and permuted P-values are reported as a measure of statistical significance. As shown for the naïve estimate, the ODA model predicted that an individual was a participant in the study if their program year cost was lower than $2,773 and a non-participant if their program year cost was equal to or greater than $2,773. The ODA model correctly classified

85.03% of participants and 71.74% of non-participants according to their cost. Classification performance was relatively strong (ESS = 56.77%) and statistically significant (*P* < 0.0001). However, after controlling for confounding via matching, ODA reported only a relatively weak clinical effect (ESS = 10.87%) that was not statistically significant (*P* < 0.076).

Taken as a whole, the statistical analysis findings obtained using ODA (Table 6) are almost perfectly consistent with findings obtained using OLS regression (Table 5), whether not-adjusting or adjusting for confounding and selection bias. For both frameworks, the naïve estimate indicates that the treatment group had statistically higher costs in the intervention period compared to the non-treated group (*P* < 0.0001, for both regression and ODA), while the adjusted estimate indicates that the treatment group had non-statistically lower costs in the intervention period compared to the non-treated group (*P* = 0.193 and *P* = 0.076, for regression and ODA, respectively).

DISCUSSION

Our results demonstrate that ODA can be combined with a matching approach as a strategy that is equally effective as conventional methods to improve causal inference in program evaluations. And while we used one-to-one matching in this particular example, the ODA algorithm can be extended to any matching design where weights are used for covariate adjustment (see for example [28,29,30,31,32,33]). It is important to note that conventional and ODA analyses may not always produce consistent results. Prior studies comparing the two methods have obtained strongly divergent findings in a wide variety of real-world data and research designs [11,12]. A good rule of thumb would be to perform the program evaluation using both conventional and ODA frameworks, and then compare the resulting treatment effect estimates. If both methods provide consistent results, then the investigator should be confident that, at the very least, the estimate is insensitive to distributional assumptions required for the OLS model, and also more likely to be a reflection of the true treatment effect estimate. However, if the approaches result in conflicting treatment effect estimates, the investigator should consider the ODA derived estimate to be more robust, given that ODA uses permutation *P* values that require no distributional assumptions and are always valid.

ODA is an appealing alternative framework in program evaluation because it holds several advantages over conventional methods for assessing covariate balance, outcomes, or both, in observational studies. First, 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 (e.g., various treatment conditions or various doses of a particular treatment), 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 conventional statistics alone (for example, as is evident in the current data by the large standard errors for the treatment effect estimates presented in Table 5).

Second, within the proposed treatment effects framework, ODA can also help explain (a) how individuals self-select in observational studies (by identifying group membership based on the cut-point on any given covariate) [10], and (b) how individuals are likely to respond to the intervention (by identifying where individuals are relative to the cutpoint on the outcome) [34].

Such detail can allow administrators to fine-tune the enrollment criteria to target those individuals who will most likely benefit from the program [20], while concomitantly allowing administrators to improve their estimates of which individuals actually benefit from the program.

Finally, ODA can be implemented using cross-validation to assess the generalizability of the model to new candidates for participation in the existing intervention, or to initiate the intervention in other settings. Cross-validation is less important if the goal is only to estimate treatment effects of the intervention [19,20].

While this paper specifically focused on creating a framework in which machine learning and matching approaches can be combined to improve causal inference in program evaluation, there are several additional ways in which machine learning techniques can be applied in causal inferential work. For example, Linden and Yarnold [20] use classification tree analysis (CTA) to characterize the nature of individuals who choose to participate in observational studies, while Athey & Imbens [35] modify the conventional classification and regression trees (CART) approach to estimate heterogeneous causal effects in such studies. CTA has also been proposed as an approach to identify potential instrumental variables (IV) that may provide an unbiased estimate of the causal effect of intervention on the outcome [10]. An IV is a variable that is correlated with the intervention, but not associated with unobserved confounders of the outcome [36]. Similarly, CTA can be used to identify causal mediation effects. A mediator is an intermediate variable which lies on the casual pathway between treatment and outcome [37]. A CTA model would be generated to predict the outcome, forcing the inclusion of the mediator after the treatment (to ensure correct temporal alignment), as well as including other covariates to control for confounding. In such a model, the extent of mediation effects can be elucidated by assessing the ESS and *P*-values for each node along the pathway from treatment to outcome via the mediator. As indicated by these examples, the application of machine-learning techniques to improve causal inference in observational studies is open to much further exploration. And 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 [38], and the development of specific sensitivity analyses for these applications [39] to ensure that the resulting models remain robust to changes in assumptions and inputs.

In summary, when ODA is combined with matching techniques within a treatment effects framework, the results are consistent with conventional approaches. However, given that ODA provides additional dimensions and robustness to the analysis are available than what can currently be achieved using conventional approaches, it offers an appealing alternative. More broadly, health researchers should consider the many potential uses of machine learning algorithms to improve causal inference in observational studies.

REFERENCES

1. Campbell, D. T., & Stanley, J. C. (1966) *Experimental and Quasi-Experimental Designs for Research*. Chicago, IL: Rand McNally.
2. Linden, A. (2007) Estimating the effect of regression to the mean in health management programs. *Disease Management and Health Outcomes*, 15, 7-12.
3. Linden, A. (2013) Assessing regression to the mean effects in health care initiatives. *BMC Medical Research Methodology,* 13, 1-7.
4. Stuart, E.A. (2010) Matching methods for causal inference: a review and a look forward. *Statistical Science,* 25, 1–21.
5. Rubin, D. (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.
6. Linden, A., & Roberts, N. (2004) Disease management interventions: What's in the black box? *Disease Management*, 7, 275-291.
7. Linden, A., Butterworth, S., & Roberts, N. (2006) Disease management interventions II: What else is in the black box? *Disease Management*, 9, 73-85.
8. 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.
9. Linden, A. (2015) Graphical displays for assessing covariate balance in matching studies. *Journal of Evaluation in Clinical Practice,* 21, 242–247.

10. Linden, A., & Yarnold, P. R. (*In Print*) Using machine learning to assess covariate balance in matching studies. *Journal of Evaluation in Clinical Practice*

11. Yarnold, P.R., & Soltysik, R.C. (2005) *Optimal data analysis: A Guidebook with Software for Windows* Washington, DC: APA Books.

12. Yarnold, P.R., & Soltysik, R.C. (2016) *Maximizing Predictive Accuracy*. Chicago, IL: ODA Books. DOI: 10.13140/RG.2.1.1368.3286

13. Yarnold, P.R., & Soltysik, R.C. (1991). Theoretical distributions of optima for univariate discrimination of random data. *Decision Sciences, 22*, 739-752.

14. Grimm, L.G., & Yarnold, P.R. (Eds.). *Reading and Understanding Multivariate Statistics*. Washington, D.C.: APA Books, 1995.

15. Grimm, L.G., & Yarnold, P.R. (Eds.). *Reading and Understanding More Multivariate Statistics*. Washington, D.C.: APA Books, 2000.

16. 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.

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

18. 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.

19. 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.

20. Linden, A., & Yarnold, P.R. (*In Print*) Using data mining techniques to characterize participation in observational studies. *Journal of Evaluation in Clinical Practice.*

21. Linden, A. (2011) Identifying spin in health management evaluations. *Journal of Evaluation in Clinical Practice,* 17, 1223-1230.

22. Rubin, D.B. (2001) Using propensity scores to help design observational studies: application to the tobacco litigation. *Health Services & Outcomes Research Methodology*, 2, 169–188.

23. Rosenbaum, P.R., & Rubin, D.B. (1983) The central role of propensity score in observational studies for causal effects. *Biometrika,* 70, 41–55.

24. Rosenbaum, P.R. (1989) Optimal matching for observational studies. *Journal of the American Statistical Association,* 84, 1024-1302.

25. Flury, B.K. & Reidwyl, H. (1986) Standard distance in univariate and multivariate analysis. *The American Statistician,* 40, 249–251.

26. Huber, P.J. (1967) The behavior of maximum likelihood estimates under nonstandard conditions. In *Vol. 1 of Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, 221–233. Berkeley: University of California Press.

27. White, H.L., Jr. (1980) A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*, 48, 817–838.

28. Linden, A., & Adams, J.L. (2010) Using propensity score-based weighting in the evaluation of health management programme effectiveness. *Journal of Evaluation in Clinical Practice, 16,* 175-179.
29. Linden, A., & Adams, J.L. (2010) Evaluating health management programmes over time. Application of propensity score-based weighting to longitudinal data. *Journal of Evaluation in Clinical Practice, 16,* 180-185.
30. 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.
31. 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.
32. 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.
33. 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.
34. Linden, A., Yarnold, P.R., & Nallomothu, B.K. (*In Print*) Using machine learning to model dose-response relationships. *Journal of Evaluation in Clinical Practice*

35. Athey, S., & Imbens, G. (2015) Recursive partitioning for heterogeneous causal effects. *Working Paper*. Downloadable from http://arxiv.org/abs/1504.01132 [Accessed on 14 May 2016].

36. Linden, A., & Adams, J. (2006) Evaluating disease management program effectiveness: an introduction to instrumental variables. *Journal of Evaluation in Clinical Practice*, 12, 148-154.

37. Linden, A., & Karlson, K.B. (2013) Using mediation analysis to identify causal mechanisms in disease management interventions. *Health Services and Outcomes Research Methodology*, 13, 86-108.

38. Linden, A., Adams, J., & Roberts, N. (2004) Evaluating disease management program effectiveness: An introduction to survival analysis. *Disease Management*, 7, 180-190.

39. 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 [10,21]). 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 (from [10,21]). 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 [10,21]). 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 (from [10,21]). Values represent cut-points on the covariate, and values in parentheses represent sensitivity (for participants) and specificity (for non-participants).

			Effect	
	Participants	Matched Controls	Strength	
	(*N*=276)	(*N*=276)	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

Table 5: Program year costs for participants and non-participants (unadjusted), and for matched pairs as estimated by ordinary least squares (OLS) regression. Values are reported as mean (standard error).

	Participants	Non-Participants	Difference	*P*-value	95% CI
Naïve estimate (unadjusted)[1]	7,325 (513)	3,287 (246)	4,038 (569)	< 0.0001	2,922, 5,154
Matched pairs OLS regression[2]	5,709 (449)	7,210 (1,060)	-1501 (1,151)	0.193	-3,762, 760

[1] 374 participants, 1628 non-participants

[2] 274 matched pairs

Table 6: Program year costs for participants and non-participants (unadjusted), and for matched pairs as estimated by optimal discriminant analysis (ODA). Values are reported as cut-points on program year costs, and values in parentheses represent sensitivity (for participants) and specificity (for non-participants). Permuted *P*-values are derived using 25,000 Monte-Carlo experiments.

	Participants	Non-Participants	Effect Strength Sensitivity	*P*-value
Naïve estimate (unadjusted)[1]	< 2,773 (85.03%)	≥ 2,773 (71.74%)	56.77%	< 0.0001
Matched pairs ODA[2]	< 4,629 (49.64%)	≥ 4,629 (61.23%)	10.87%	0.076

[1] 374 participants, 1628 non-participants

[2] 274 matched pairs