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

Journal of Evaluation in Clinical Practice ISSN 1356-1294

Using propensity score-based weighting in the evaluation of health management programme effectiveness

Ariel Linden DrPH MS[1] and John L. Adams PhD[2]

[1]President, Linden Consulting Group, Hillsboro, OR USA
[2]Senior Statistician, RAND Corporation, Santa Monica, CA, USA

Keywords
health management programmes, inverse probability of treatment weights, propensity score

Correspondence
Ariel Linden
Linden Consulting Group
Hillsboro OR 97124
USA
E-mail: alinden@lindenconsulting.org

Accepted for publication: 19 February 2009

doi:10.1111/j.1365-2753.2009.01219.x

Abstract

When the randomized controlled trial is unfeasible, programme evaluators attempt to emulate the randomization process in observational studies by creating a control group that is essentially equivalent to the treatment group on known characteristics and trust that the remaining unknown characteristics are inconsequential and will not bias the results. In recent years, adjustment procedures based on the propensity score, such as matching and subclassification, have become increasingly popular. A new technique that has particular appeal for evaluating health management programmes uses the propensity score to create a weight based on the subject's inverse probability of receiving treatment. This weighting mechanism removes imbalances of pre-intervention characteristics between treated and non-treated individuals, and is then used within a regression framework to provide unbiased estimates of treatment effects. This paper presents a non-technical introduction of this technique by illustrating its implementation with data from a recent study estimating the impact of a motivational interviewing-based health coaching on patient activation measure scores in a chronically ill group of individuals. Because of its relative simplicity and tremendous utility, propensity-score weighting adjustment should be considered as an alternative procedure for use with observational data to evaluate health management programme effectiveness.

Introduction

The simplest explanation for why we conduct evaluations is to determine if an intervention is effective in achieving a given outcome. While superficially this appears to be a basic 'cause and effect' formula, in fact there are many components that need to be accounted for. The randomized controlled trial (RCT) has always been considered the gold standard in study designs because as its name implies, individuals are randomly assigned to receive either treatment or control, thereby giving each person an equal probability to be chosen for the intervention. This procedure is intended to ensure that individuals assigned to either group are comparable on both known and unknown characteristics, and thus unconditionally exchangeable. Any differences found in outcome measures between the study groups can then be attributed to the programme intervention and not biased by baseline differences in group characteristics or an effect of confounders.

While the RCT remains the primary choice of design for inferring a causal relationship between the intervention and outcome, the commercial application of health management programmes generally precludes the use of this study design because purchasers commonly believe that all individuals meeting programme eligibility will benefit from it. Thus, programmes specifically target for enrolment all individuals classified as high 'risk' (in wellness programmes, *risk* may indicate the prospect of developing an illness, while disease management programmes typically qualify *risk* as the likelihood of incurring high medical costs in the near future). It is clear that selection bias is a threat to validity using such an enrolment strategy and that the true outcome can be further obscured by regression to the mean (given that a large subset of those initially classified as high risk are naturally likely to appear as lower risk following the programme intervention) [1].

When the RCT is unfeasible, programme evaluators attempt to emulate the randomization process in observational studies by creating a control group that is essentially equivalent to the treatment group on known characteristics and trust that the remaining unknown characteristics are inconsequential and will not bias the results [2]. In many disciplines, conventional regression modelling remains the most common approach used to account for pre-intervention differences between groups, even though there is sufficient evidence that these methods may provide biased results, most notably in the presence of time-dependent confounders [3,4].

In recent years, however, adjustment techniques based on the propensity score have become increasingly popular. The

propensity score, defined as the probability of assignment to the treatment group conditional on covariates [5], controls for pre-intervention differences between enrolled and non-enrolled groups. Propensity scores can be derived from a logistic regression equation that reduces each participant's set of covariates to a single score. It has been demonstrated that, conditional on this score, all observed pretreatment covariates can be considered independent of group assignment, and in large samples, covariates will be distributed equally in both groups and will not confound estimated treatment effects [5].

Once the propensity score has been estimated in a given dataset, treatment effects can then be modelled. Matching treated to non-treated individuals on their propensity score [6,7,8] appears to be the most popular propensity scoring technique used in evaluating health management programmes [9]. However, matching on the propensity score has some inherent limitations when evaluating such programmes. Health management interventions generally include a small number of active participants contrasted with a very large number of non-participants. Successful matches are generally found for all programme participants, leaving most non-participants in the population unmatched and thereby excluded from the analysis. There are two consequences of this: (1) treatment effects many be statistically insignificant because of the confluence of a small treatment group and a rare-event outcome (e.g. hospital admissions, emergency department visits); and (2) by excluding data from the unmatched population, the effect of non-treatment in the remaining population with the disease is not captured. Thus, we gain no insight as to how well the programme chose its participants, or if the programme could have been effective on those individuals not explicitly targeted for the intervention [10].

Stratification is another propensity score adjustment approach. Outcomes are arranged into quintiles based on the range of propensity scores divided into treated and non-treated groups. This allows the evaluator to review outcomes between groups at each stratum, as well as to observe differences within groups between strata. This can be done using statistical tests, but often visual inspection alone can highlight important effects [10]. It has been shown that stratification of the propensity score into quintiles (generally referred to as subclassification) can remove over 90% of the initial bias owing to the covariates used to create the propensity score [11,12]. If important within-subclass differences between cohorts are found on some covariates, it could be concluded that the covariate distributions did not overlap sufficiently to allow subclassification to adjust for these covariates, raising concern about the model's ability to draw valid conclusions about the results. In such cases, alternate analytic adjustments should be considered [12].

A recent addition to the inventory of propensity score-based adjustment procedures uses weighted regression to estimate the effect of treatment on an outcome. The weight used in the regression is based on the conditional probability of an individual receiving his or her own treatment. More specifically, participants have a weight equal to the inverse of the estimated propensity score (1/propensity score), and non-participants have a weight equal to the inverse of 1 minus the estimated propensity score (1/1 – propensity score). This weighting scheme, called the 'inverse probability of treatment weights' (IPTW) [4,13] adjusts for pre-intervention differences between participants and non-participants. The IPTW can then be used in almost any type of regression model, for either point-treatment or longitudinal studies. Used in the longitudinal context, these regression models are generally referred to as marginal structural models (MSM) [13].

This paper presents a non-technical introduction to the IPTW as an alternative propensity score-based approach to providing an unbiased estimate of a health management programme's treatment effect. We will start by describing the mechanism by which IPTW achieves balance between treated and non-treated groups followed by an illustration of how the IPTW is then used in weighted regression models. We illustrate the implementation of this technique with data from a recent study estimating the impact of a motivational interviewing-based health coaching on patient activation measure (PAM) [14] scores in a chronically ill group of individuals [15]. We then provide a discussion of limitations of the technique.

Inverse probability of treatment weighting

The concept of inverse probability of selection weighting originated in the survey sciences over 50 years ago to adjust for imbalances in sampling pools [16] and continue to be regularly used in complex survey designs. As an example, the National Health and Nutrition Examination Survey (NHANES) [17], which provides a snapshot of the health and nutrition status of the US population, purposely over-samples certain subgroups such as non-Hispanic blacks, Mexican Americans and persons age 12–19 years. Table 1 illustrates that after applying the inverse probability of selection weights, the NHANES sample population has a similar distribution to that of the non-institutionalized US population. Stated another way, the weighting mechanism standardized the sample to the greater population from where they were drawn, thereby allowing unbiased population estimates to be calculated. For example, being herpes positive, is correlated with non-Hispanic black race/ethnicity. The weighted prevalence estimate of 17.9% was much lower than the over-sampled unweighted estimate of 24.1% [18].

Robins [13] and Robins *et al.* [4] have applied this weighting concept to the study of treatment effects in observational studies. Here, the weight is the inverse probability of the subject's treatment status (where the probability of receiving treatment is the estimated propensity score), thus treated subjects are given a

Table 1 Race/ethnicity distribution (percent) of the US non-institutionalized population (2000 census) and the 1999–2002 National Health and Nutrition Examination Survey (NHANES) interview sample [17,18]

	NHANES		
	US population	Un-weighted	Weighted
Non-Hispanic Black	13	25	12*
Non-Hispanic White/other	78	47	n/a†
Mexican American	9	28	9
12- to 19-year-olds	12	24	12

*Numbers vary slightly because of rounding.
†Data not available.

weight of 1/(propensity score) and non-treated subjects are given a weight of 1/(1 – propensity score) [4].

This weighting mechanism can be thought of as creating a pseudo-population comprised of 'copies' of the original subjects who account not only for themselves but for subjects with similar characteristics who received the alternate exposure [19]. More specifically, a programme participant with a low estimated propensity score will contribute many more copies of himself or herself to the pseudo-population than a participant with a high estimated propensity score (for example, a programme participant with a propensity score of 0.01 will contribute 100 copies while a participant with a propensity score of 1.0 will contribute only one copy). The interpretation of the weighting formula for non-participants is analogous.

As a result of removing any existing association between pre-intervention covariates and treatment, the IPTW allows us to view the pseudo-population as one in which all individuals are considered conditionally exchangeable. Thus, the IPTW has a two-pronged effect: (1) it ensures that balance is achieved between the treated and non-treated groups on pre-intervention characteristics [20]; and (2) provides us with greater confidence that treatment effect estimates derived from observational data are unbiased (presuming that all sources of bias were accounted for in the estimated propensity score) [4]. In effect the IPTW weights the analysis so it looks as much as possible like an RCT.

Model estimation using IPTW

Unbiased treatment effects can be estimated by fitting the appropriate regression model using the IPTW as the specified weight (for example, in the Stata software package one would specify the IPTW as either an analytic weight or sampling weight). Like any other outcome variable, the choice of regression model depends on the distribution of the outcome variable. This can be logistic regression for dichotomous variables, ordinary least squares (OLS) for continuous variables, Poisson for rates or rare events, and Cox regression for survival or censored cases. Some researchers prefer the use of generalized linear modelling for its flexibile distributional assumptions [21].

Regardless of the choice of model, the usual standard errors generated by the weighted model will tend to be mis-specified, which in turn will produce mis-specified confidence intervals and potentially invalidate tests of significance. This issue can be circumvented via the use of robust standard errors [22] or by bootstrapping [23] the beta coefficient of the treatment parameter. These procedures are commonly available in most statistical or econometrics software.

As the IPTW implicitly incorporates covariates into the statistical model by way of the propensity score, it is generally sufficient to include only the treatment variable into the structural component of the regression (e.g. the right-side). However, other meaningful covariates can be added as well. In fact, Robins and his associates [24,25,26] have recently introduced the notion of 'doubly robust' (DR) estimators, which requires the use of IPTW and covariates within the same regression model. In a causal inference model, an estimator is DR if it remains consistent when either model (propensity score or outcomes regression) is correctly specified. Therefore, an evaluator is given two chances, instead of only one, to make a valid inference. Emsley *et al*. [27] provide a very approachable description of the DR estimator, and tutorial on its implementation in the Stata software package.

Example of the IPTW technique to evaluate a health coaching programme

Our data comes from a recent study which evaluated the impact of MI-based health coaching on a chronically ill group of individuals [15]. Measures were chosen that could be directly attributed to a health coaching intervention on chronic illness: self-efficacy for managing chronic illness [28], patient activation measure [14], stage of readiness to change [29], lifestyle change and perceived global health status using the EQ-5D visual analogue scale [30]. It was hypothesized that programme participants ($n = 106$) would show significant improvement in these measures compared with non-participants ($n = 230$). Here, we illustrate the implementation of the IPTW technique focusing only on the programme effect's on PAM scores. We encourage readers to review the companion article in this issue [15] for a more contextual and comprehensive discussion about the programme and its impact on all other outcomes.

In commercial health management programmes, there is a high probability that participants will have different characteristics than non-participants because of the specific enrolment strategy targeting those individuals who appear to be at higher risk. The unweighted pre-intervention characteristics shown in Table 2 prove this to be the case. Compared with non-participants, new programme enrolees had significantly lower: self-efficacy in managing their chronic illnesses, activation (knowledge, skills, beliefs and confidence) to partner with their health provider to manage their health and perceived overall health status. Given the treatment group's lower starting values, regression to the mean becomes a genuine threat to the validity of study outcomes. As shown in Table 2, these imbalances between participant and

Table 2 Comparison of various unweighted and weighted pre-intervention characteristics between programme participants and non-participants (adapted from: Linden A *et al.* [15])

	Unweighted*		Weighted*	
Variable	Participants	Non-participants	Participants	Non-participants
Self-efficacy (0–10)	7.2 (6.8, 7.5)	8.5 (8.3, 8.6)	7.9 (7.6, 8.2)	8.0 (7.8, 8.2)
Patient activation measure (0–100)	68.1 (65.4, 70.9)	76.6 (74.7, 78.4)	71.8 (69.0, 74.6)	74.6 (72.8, 76.6)
Perceived Health Status (0–100)	71.7 (68.9, 74.6)	79.1 (77.5, 80.7)	75.7 (73.0, 78.3)	77.2 (75.5, 78.9)

*Values are means and 95% confidence intervals (in parentheses).

non-participant groups were removed upon adjusting these variables using the IPTW.

The PAM score variable used in the outcome model was calculated using a differences-in-differences estimator. That is, the treatment effect was modelled by estimating the difference between outcome scores in the second survey period minus the baseline score (first survey) for both participants and non-participants, and then compared the difference between the two groups. An OLS regression modelled the PAM difference score using the treatment variable as a covariate and applying the IPTW as weights. Robust standard errors were generated to produce conservative confidence intervals for the weighted regression model estimates. The causal model estimated that programme participation led to improved PAM scores by an average of 4.57 points (95% confidence interval: 0.63–8.25) on a scale that ranges from 0 to 100 points.

Limitations of the IPTW technique

As with any evaluation of observational data, the foremost limitation is that we presume that all biases and confounding have been adjusted for in the model, an assumption that cannot truly be tested outside of a randomized study. One limitation specific to inverse probability weighting is that propensity scores for programme participants must be different from zero [19]. In effect, no treatment effect can be estimated for people who have no probability of receiving the treatment. In this case the study population can be redefined to a subset of the population where the probability of treatment is greater than zero. Another limitation of IPTW is that it can perform poorly when the weights for few subjects are very large. In this situation the standard errors of the treatment effect variable may underestimate the true difference between the weighted estimator and the population parameter it estimates [31]. Given these limitations, the evaluation is best served by a close inspection of these individual instances to assess whether they represent real cases or suspect data.

Conclusion

In this paper we have presented an alternative propensity score-based adjustment procedure that uses weighted regression to estimate the effect of treatment on an outcome. This method standardizes the outcome across the entire population and therefore provides treatment effect estimates had the intervention been given to everyone. Although beyond the scope of this paper, the IPTW technique has particular appeal over standard analytic adjustments if the evaluation's focus is on longitudinal data [32]. Perhaps the most salient element of the inverse probability weighting concept is its ability to correct for imbalances in pre-intervention characteristics between treated and non-treated groups across all levels or strata. This provides us with confidence that the groups are essentially equivalent (assuming there is no residual confounding), thereby allowing us to make causal inferences about treatment effects. Because of its relative simplicity and tremendous utility, propensity-score weighting adjustment should be considered as an alternative procedure for use with observational data to evaluate health management programme effectiveness.

References

1. Linden, A. (2007) Estimating the effect of regression to the mean in health management programs. *Disease Management and Health Outcome*, 15 (1), 7–12.
2. 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.
3. Freedman, D. (1999) From association to causation: some remarks on the history of statistics. *Statistical Science: A Review Journal of the Institute of Mathematical Statistics*, 14, 243–258.
4. Robins, J. M., Hernan, M. A. & Brumback, B. (2000) Marginal structural models and causal inference in epidemiology. *Epidemiology*, 11, 550–560.
5. Rosenbaum, P. R. & Rubin, D. B. (1983) The central role of propensity score in observational studies for causal effects. *Biometrika*, 70, 41–55.
6. Heckman, J., Ichimura, J. & Todd, P. (1997) Matching as an econometric evaluation estimator: evidence from evaluating a job training programme. *The Review of Economic Studies*, 64, 605–654.
7. 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.
8. Rubin, D. B. & Thomas, N. (1996) Matching using estimated propensity scores: relating theory to practice. *Biometrics*, 52, 249–264.
9. Linden, A., Adams, J. & Roberts, N. (2005) Using propensity scores to construct comparable control groups for disease management program evaluation. *Disease Management and Health Outcomes*, 13 (2), 107–127.
10. 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 (5), 914–918.
11. Cochran, W. G. (1968) The effectiveness of adjustment by subclassification in removing bias in observational studies. *Biometrics*, 24, 205–213.
12. 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.
13. Robins, J. M. (1998) Marginal structural models. In 1997 Proceedings of the Section on Bayesian Statistical Science, pp. 1–10. Alexandria, VA: American Statistical Association.
14. Hibbard, J. H., Stockard, J., Mahoney, E. R. & Tusler, M. (2004) Development of the Patient Activation Measure (PAM): conceptualizing and measuring activation in patients and consumers. *Health Services Research*, 39 (4), 1105–1026.
15. Linden, A., Butterworth, S. W. & Prochaska, J. O. (2010) Motivational interviewing-based health coaching as a chronic care intervention. *Journal of Evaluation in Clinical Practice*, 16, 166–174.
16. Horvitz, D. G. & Thompson, D. J. (1952) A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47, 663–685.
17. National Center for Health Statistics. (2009) *National Health and Nutrition Examination Survey*. Available at: http://www.cdc.gov/nchs/nhanes.htm (last accessed 16 February 2009).
18. National Center for Health Statistics. *National Health and Nutrition Examination Survey. Specifying Weighting Parameters*. Available at: http://www.cdc.gov/nchs/tutorials/Nhanes/SurveyDesign/Weighting/intro.htm (last accessed 16 February 2009).
19. Hernán, M. A., Hernández-Diaz, S. & Robins, J. M. (2004) A structural approach to selection bias. *Epidemiology*, 15, 615–625.
20. Rosenbaum, P. R. (1987) Model-based direct adjustment. *Journal of the American Statistical Association*, 82, 387–394.

21. McCullagh, P. & Nelder, A. (1989) Generalized Linear Models, 2nd Edn. London: Chapman and Hall.
22. White, H. A. (1980) A heteroscedasticity-consistent covariance matrix estimator and a direct test of heteroscedasticity. *Econometrica*, 48, 817–838.
23. Linden, A., Adams, J. & Roberts, N. (2005) Evaluating disease management program effectiveness: an introduction to the bootstrap technique. *Disease Management and Health Outcomes*, 13 (3), 159–167.
24. Scharfstein, D. O., Rotnitzky, A. & Robins, J. M. (1999) Adjusting for nonignorable drop-out using semiparametric nonresponse models. *Journal of the American Statistical Association*, 94, 1096–1120 (with Rejoinder, 1135–1146).
25. Robins, J. M. (2000) Robust estimation in sequentially ignorable missing data and causal inference models. In Proceedings of the American Statistical Association Section on Bayestan Statistical Science, pp 6–10. Alexandria, VA: American Statistical Association.
26. Bang, H. & Robins, J. M. (2005) Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61, 962–973.
27. Emsley, R., Lunt, M. & Pickles, A. (2008) Implementing double-robust estimators of causal effects. *Stata Journal*, 8 (3), 334–353.
28. Lorig, K., Chastain, R. L. & Ung, E., *et al.* (1989) Development and evaluation of a scale to measure perceived self-efficacy in people with arthritis. *Arthritis Rheum*, 32 (1), 37–44.
29. Prochaska, J. O. (1979) Systems of Psychotherapy: A Transtheoretical Analysis. Homewood, IL: Dorsey Press.
30. Shaw, J. W., Johnson, J. A. & Coons, S. J. U. S. (2005) Valuation of the EQ-5D health states development and testing of the D1 valuation model. *Medicine Care*, 43 (3), 203–220.
31. Kurth, T., Walker, A. M., Glynn, R. J., Chan, K. A., Gaziano, J. M., Berger, K. & Robins, J. M. (2006) Results of multivariable logistic regression, propensity matching, propensity adjustment, and propensity-based weighting under conditions of nonuniform effect. *American Journal of Epidemiology*, 163 (3), 262–270.
32. Hernán, M. A., Brumback, B. & Robins, J. M. (2000) Marginal structural models to estimate the causal effect of zidovudine on the survival of HIV-positive men. *Epidemiology*, 11, 561–570.

