Evaluation & the Health Professions
34(2) 164-180

DOI: 10.1177/0163278710376824
http://ehp.sagepub.com

Designing a Prospective Study When Randomization is Not Feasible

Ariel Linden[1]

Abstract

When conducting a randomized controlled trial (RCT) is unfeasible, the goal is to replicate the randomization process by creating a control group that is essentially equivalent to the treatment group on known pre-intervention characteristics and assume that the remaining unknown characteristics will not bias the results. The strategies proposed in this article are based on the thesis that since *only* pre-intervention characteristics are used for adjustment, a comparable control group can be established as soon as the participant group is identified. Consequently, outcomes can be observed immediately after launching the initiative rather than waiting until study completion. The benefit is that significant treatment effects can be observed as they occur, or alternatively, the initiative can be cancelled if treatment effects are not attained by a certain time point. Although these methods can never ensure the same level of validity as in an RCT, they are considered robust alternatives when randomization is impractical, and therefore a compelling study design for many commercial initiatives, such as disease management programs, benefit design changes,

[1] Linden Consulting Group, OR

Corresponding Author:
Ariel Linden, 6208 NE Chestnut Street, Hillsboro, OR 97124, USA
Email: alinden@lindenconsulting.org

and pay-for-performance efforts. An obvious constraint is that treated participants must first be identified before suitable controls can be found. The preferred strategy is to enroll the entire treatment group within a narrow time frame. An alternative option is to have periodic enrollment periods with their respective treatment and control cohorts. The concept proposed in this article is intended to offer a robust alternative to the inadequate strategies currently being used in many health care settings where study findings may not be trusted, and thus decision makers remain uninformed as to whether an initiative is worth continuing or cancelled.

Keywords
propensity score, matching, weighting, stratification, covariate balance

Introduction

In the commercial health care industry, programs, pilots, and other initiatives are often implemented without a concrete, detailed evaluation plan. This becomes problematic when the evaluation is ultimately conducted because sample size may be insufficient to detect a program effect, data from a suitable control group may not be readily available, or measures required for mitigating bias (e.g., demographic data, comorbidities, etc.) may not have been collected. In such a situation, the evaluation will be of insufficient rigor to provide meaningful information to decision makers as to whether the program is worth continuing. As a result, ineffective pilots may be expanded to full-fledged programs or effective initiatives may be terminated. Thus, organizations are well served by spending time upfront to design a thoughtful evaluation plan.

Two common misconceptions about evaluation design are that: (a) if a randomized controlled trial (RCT) cannot be implemented, the remaining options are all problematic and similarly flawed, and (b) evaluation activities must take place at the conclusion of the intervention to draw out meaningful information. In fact, when the RCT is unfeasible there is an array of robust techniques available, which attempt to replicate the randomization process by creating a control group that is essentially equivalent to the treatment group on known pre-intervention characteristics (and assuming that the remaining unknown characteristics will not bias the results) (Rubin, 2007). Furthermore, once this robust evaluation strategy is developed, feedback can be provided to decision makers at any point during the intervention, allowing them to make timely decisions regarding

the continuation or modification of the initiative. This article provides a framework and describes at a high level the techniques for following such an approach.

A Conceptual Framework

The conceptual framework proposed in this article is based on the thesis that since *only* pre-intervention characteristics are used for adjustment, a comparable control group can be established as soon as the participant group is identified. Consequently, outcomes can be observed immediately after launching the initiative rather than waiting until study completion. This involves a three-step process. First, as soon as the participant group is established, a suitable method is used to construct a control group. Next, tests are conducted to ensure that the groups are comparable, that is balance between groups has been achieved on all pre-intervention variables. Finally, the outcome measure is tracked prospectively, comparing the treated to nontreated groups using the appropriate statistics. Each of these steps will be described in turn.

Creating a Comparable Control Group Based on the Propensity Score

Once participants have been selected and a pool of potential controls identified, the first step is to determine which controls to select, and then if needed adjust for any baseline differences between the intervention and control groups. In many disciplines, conventional regression modeling remains the most common approach used to account for pre-intervention differences between groups. However, there is sufficient evidence that these methods may provide biased results, most notably in the presence of time-dependent confounders (Freedman, 1999; Robins, Hernán, & Brumback, 2000). As a result, researchers have sought to develop more robust adjustment methods to make treatment and control groups comparable.

In recent years, 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 observed covariates (Rosenbaum & Rubin, 1983), controls for pre-intervention differences between treated and nontreated groups. Propensity scores are generally 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 (Rosenbaum & Rubin, 1983).

Once the propensity score has been estimated in a given dataset, treatment effects can then be modeled. Matching treated to nontreated individuals on their propensity score (Dehejia & Wahba, 1999; Heckman, Ichimura, & Todd, 1997; Linden, Adams, & Roberts, 2005; Rubin & Thomas, 1996) is perhaps the most popular technique, and there are several different matching algorithms currently in use, including pair-wise matching (also called one-to-one matching), matching using propensity score categories (Dehejia & Wahba, 1999), matching based on the Mahalanobis distance (Rubin, 1980) and kernel density matching (Heckman et al., 1997).

Stratification is another propensity score adjustment approach. Outcomes are generally arranged into quintiles based on the range of propensity scores divided into treated and nontreated groups. This allows the evaluator to analyze outcomes between groups within each stratum, as well as to observe overall differences between groups across all strata (Linden & Adams, 2008). 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 due to the covariates used to create the propensity score (Cochran, 1968; Rosenbaum & Rubin, 1984).

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 most commonly used weighting scheme is the "inverse probability of treatment weights" (Robins et al., 2000), which is intended to provide an estimate of the average treatment effect (ATE) in the population for which treatment is appropriate (Imai, King, & Stuart, 2008). Participants receive a weight equal to the inverse of the estimated propensity score (1/propensity score) and nonparticipants have a weight equal to the inverse of 1 minus the estimated propensity score (1/(1 − propensity score)). However, the evaluator may be more interested in setting the distribution of covariates to be equal to that of the treated subjects and then estimating the average treatment effect on the treated (ATT), where the quantity of interest is to be the treatment effect averaged over only the treated units (Imai et al., 2008; Imbens, 2004). In this case, participants are given a weight of 1 and nonparticipants are given a weight of the (propensity score)/(1 − propensity score; Nichols, 2008). This ATT weighting mechanism makes the control group's outcomes represent the counterfactual outcomes of the treatment group by making the two groups similar with respect to observable pre-intervention characteristics (those variables

included in the propensity score model; Nichols, 2008). Once the weights are constructed, they can then be used within the regression model framework for either point-treatment (Linden & Adams, 2010a) or longitudinal studies (Linden & Adams, 2010b). The choice of which weight to use depends on what question the evaluator seeks to answer.

Testing Covariate Balance

Fundamental to any study, whether randomized or observational, is the requirement that treatment and control groups be comparable on pre-intervention characteristics. Imbalances in covariates between groups lead to systematic biases that will limit the validity of study findings. In the RCT, we assume that balance is naturally achieved in both observed and unobserved covariates. Due to selection bias, in observational studies we cannot make this assumption and therefore must assess covariate balance on observed characteristics alone. There are several methods available to assess covariate balance including standardized differences (Flury & Reidwyl, 1986), Kolmogorov-Smirnov equality of distributions test (Conover, 1999), or diagnostic plots such as quantile–quantile plots or box plots (Chambers, Cleveland, Kleiner, & Tukey, 1983). Austin (2009) provides a comprehensive discussion on balance checking in propensity score matching studies.

The standardized difference is perhaps the most traditionally used measure of balance and is simple to compute (Flury & Reidwyl, 1986):

$$d = \frac{100 \times \left(\bar{X}\,treatment - \bar{X}\,control\right)}{\sqrt{(s^2 treatment + s^2 control)/2}},$$

where the numerator is the difference in means between the groups and the denominator is the pooled standard deviation. The appeal of this method is that it is indifferent to the unit of measurement and insensitive to sample size. Normand et al. (2001) suggest that a standardized difference of less than 10% is indicative of good balance; however, there is no empirical evidence to support the use of any particular cutoff point. In fact, given that the standardized difference is based on Cohen's *d* statistic for effect size (Cohen, 1988), one could argue that a value <20% is a reasonable absolute cutoff.

Rosenbaum and Rubin (1985) take a slightly different approach with the standardized difference (which they refer to as standardized bias). They perform the calculation first with the treated versus the nontreated population

and then again comparing the treated versus the matched controls. However, they keep the denominator the same in both calculations (using the treated vs. the nontreated population's pooled standard deviation). Because the denominators are the same, a comparison of pre- and post-matching standardized differences shows the extent to which matching has made the means closer. For weighting methods, approaches similar to those for matching (e.g., standardized differences, diagnostic plots, box-plots, or histograms) can be used; however, use of these methods must incorporate the weights. It is not uncommon for evaluators to repeat the matching process several times or alternate between methods before achieving optimal balance.

Tracking Outcomes Prospectively

Tracking intervention outcomes is rather straightforward but differs slightly depending on the adjustment model used. In general, comparisons between treatment and control groups would be made at each observation point of the study, including outcome data (the dependent variable) from pre-intervention periods used to create the propensity score. For example, if data were collected monthly and 12 months of pre-intervention observations were used to create the propensity score, one could plot the 12 baseline observations (means, medians, or other statistic of interest) for treated and control groups and then add new monthly observations as they became available after initiation of the intervention.

Observations from propensity-scored matches are the easiest to interpret because they represent actual values (as in an RCT). Conversely, weighted observations are adjusted values, meaning that individuals within the treatment group cannot be directly compared to individuals in the nontreated group. This is not a critical issue because it is the group point estimates that are meaningful for comparison, not individual values. Observations stratified into propensity score quintiles are perhaps the most time consuming to plot but will ultimately provide the most detailed information about treatment effects across the entire population (Cochran, 1968; Rosenbaum & Rubin, 1984).

Point estimates and other statistics must be appropriate to the outcome measure under study for interpretation of results to be useful. A common mistake is to emphasize the difference in group means alone, without incorporating potential variability in the distribution of values. One can easily misinterpret a mean difference between groups as being a significant intervention effect, when in fact the confidence intervals (CI) suggest otherwise.

Another common mistake is to use the mean of the variable when the distributional characteristics indicate that using the median or transforming the variable is more appropriate (as is the case when analyzing cost data). When using a matching strategy, statistical models must be chosen to account for the paired (or dependent) nature of the data. Paired *t* tests and Wilcoxon signed rank tests are options for continuous variables, whereas McNemar's test is suitable for binary variables (Austin, 2008). For those preferring to use a regression approach, most statistical software packages generally allow for adjustment of standard errors by clustering at the matched pair level.

Statistical analyses may be performed at each time interval; however, they may be more useful as a temporal guide rather than a definitive indication of a treatment effect. Only a comprehensive evaluation (usually conducted at the completion of the study) can truly adjust for the effects of time-varying confounders, attrition, and other sources of bias. Nonetheless, meaningful statistics can be provided at each observation point in the study. For example, 95% CI can be computed for each group's point estimate or for the difference score between groups at each time period. To improve robustness, correction approaches should be considered to adjust for the inflated risk of committing a Type I error due to multiple testing (such as the Bonferroni adjustment, etc.; Hochberg & Tamhane, 1987).

Example: A Health Management Program

To demonstrate the proposed conceptual framework, we use data from a health management program. The program targets individuals with chronic conditions and includes a nurse-based intervention intended to improve clinical indices and manage health services use. The data consist of 24 monthly observations for 155 program participants and 7,713 nonparticipants (for a total of 188,832 observations). See Linden and Adams (2010b) for a more comprehensive description of the data. Here, we focus on physician office visits as the primary outcome under study.

The first 12 months of data representing the preprogram baseline period are presented in Table 1. As the standardized difference shows, participants were older, sicker, and more costly than nonparticipants. Given the obvious selection bias, adjustment is required to make the groups comparable on observed characteristics. These findings are further supported by Figures 1 and 2, which illustrate that the unadjusted monthly office visit rates for the treated group are consistently higher than the nontreated population over the entire course of the study. Figure 1, in particular, illustrates what

Table 1. Baseline (12 months) Characteristics of Program Participants, Nonparticipants, and Matched Controls

Variable[a]	Participants	Nonparticipants	Matched Controls	Std Diff[b]	Std Diff[c]
N	155	7,713	155		
Age	56.45 (9.2)	46.59 (11.0)	57.53 (10.6)	97%	11%
Female (%)	45.81	53.26	52.90	15%	14%
Congestive heart failure (%)	9.68	0.65	10.32	42%	2%
Diabetes (%)	67.1	9.87	65.81	145%	3%
Hospital admissions	0.21 (0.5)	0.04 (0.3)	0.22 (0.6)	42%	2%
Emergency department visits	0.43 (1.1)	0.12 (0.4)	0.41 (1.1)	37%	2%
Physician office visits	9.65 (6.3)	3.81 (4.4)	10.02 (8.3)	107%	5%
Prescriptions	46.45 (28.4)	11.62 (16.4)	48.34(35.8)	150%	6%
Total costs	$13,522 (17,585)	$3,107 (8,857)	$15,482 (31,700)	75%	8%

[a] Unless otherwise noted, variables presented are means and standard errors.
[b] Standardized difference (Participants minus nonparticipants).
[c] Standardized difference (Participants minus matched controls).

a decision maker would likely see in the absence of a robust study design or analytic adjustment. Visually, it is impossible to determine whether the treatment group fared better than the controls, except for the obvious spike that occurs in the first program month. Figure 2 adds some element of statistical value for the decision maker, but because the groups are so incomparable, these statistics are meaningless.

To mitigate the apparent selection bias, the propensity score was estimated for each individual in the population using all the variables listed in Table 1 (aggregated into one annual block rather than 12 monthly increments). One-to-one matching was performed using the nearest neighbor method with no replacement. As shown in Table 1, the standardized differences between treatment and matched control units indicated that very good balance was achieved on the observed covariates.

For the purpose of exposition, propensity score weighting was also performed. ATT weights were constructed as described earlier in this article. Theoretically, both methods should produce similar results, because both procedures adjust the nontreated units to make them similar to treated units, with no adjustment made to the treated units. Figure 3 illustrates the monthly office rates for the matched pairs and using the weighted adjustment. The purpose of this visual display is to illustrate that both methods performed similarly over the entire course of the study, suggesting that in

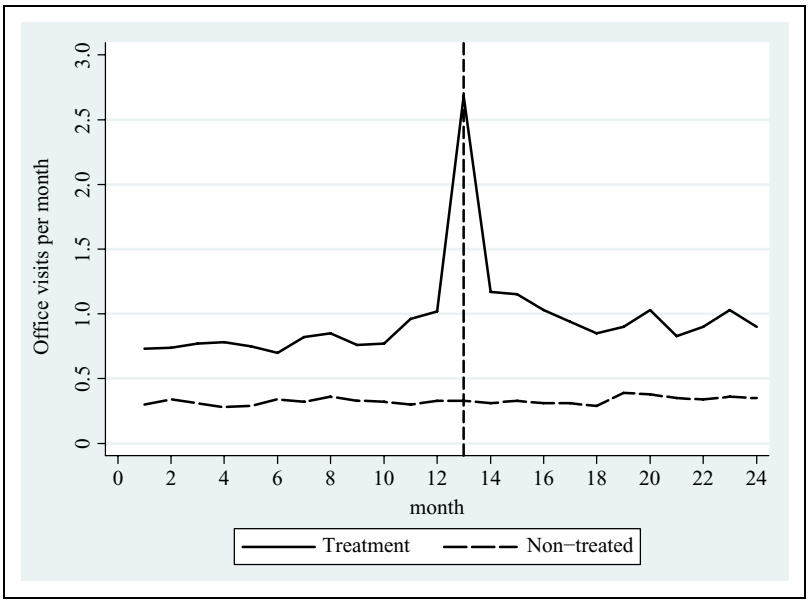

Figure 1. Unadjusted physician office visits per month for 155 program participants and 7,713 nonparticipants. The vertical dashed line represents program initiation.

this case the evaluator may feel confident choosing either adjustment strategy to achieve similar results. As described earlier, presenting point estimates alone do not tell us whether the differences between the two groups are statistically meaningful. A more appropriate visual display would incorporate statistical values such as those displayed in Figure 4. Figure 4 plots the difference (treatment effect) in monthly physician office visit rates and 95% (CI) between treated group and their propensity score matched controls (nearly identical findings were derived using the weighting model and are therefore not presented here). As illustrated, the monthly 95% (CI) cross zero for the entire baseline period, further supporting the findings in Table 1 that the groups were balanced on covariates. In the first study month (corresponding to Month 13), study participants showed a dramatic increase in office visits and they remained significantly higher than controls for the next four months (as indicated by the CI not crossing zero from Month 13 to 17). From the sixth program month (corresponding to Month 18) onward, the office visit rate returned to a level not statistically different than controls. Assuming this outcome variable is the primary

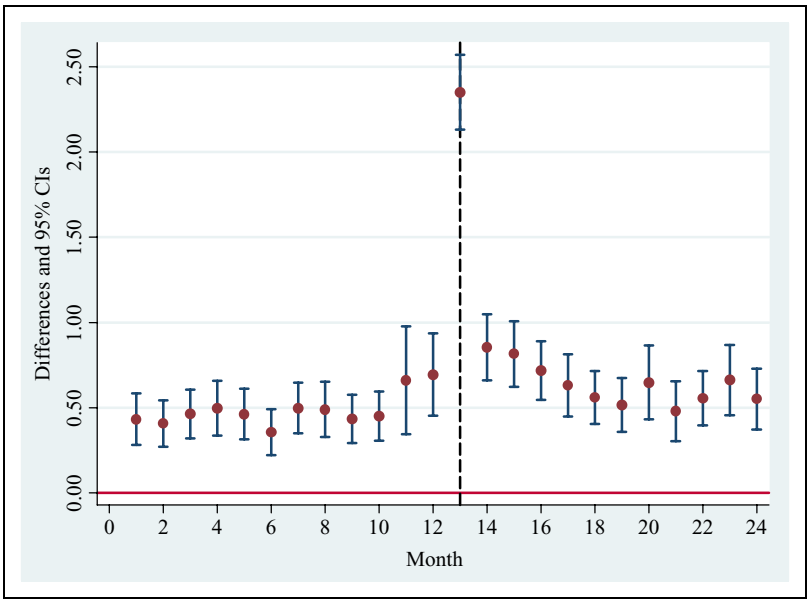

Figure 2. Naïve estimate of program effect on physician office visit rate. Values and 95% confidence intervals represent the treated group's monthly office visit rate relative to the nontreated comparison group (red line at zero). The vertical dashed line represents program initiation.

measure under study, one could conclude that the intervention begins to take effect in the first month of the program and lasts for about 6 months. In a hypothetical situation in which decision makers determined a priori that the program would be terminated as soon as there was no longer a treatment effect, Month 6 of the intervention (corresponding to Month 18 in the Figure) would have been the concluding month of the study. A more likely scenario would be to allow the program to run over the course of 12 months (as currently plotted) and then perform a more comprehensive program evaluation at the end of that period.

Additional Considerations

There are two additional factors that should be considered to maximize the effectiveness of the study framework presented here. First, every effort must be made to enroll all the individuals assigned to the treatment arm as soon as possible. As a practical matter, the propensity score estimation

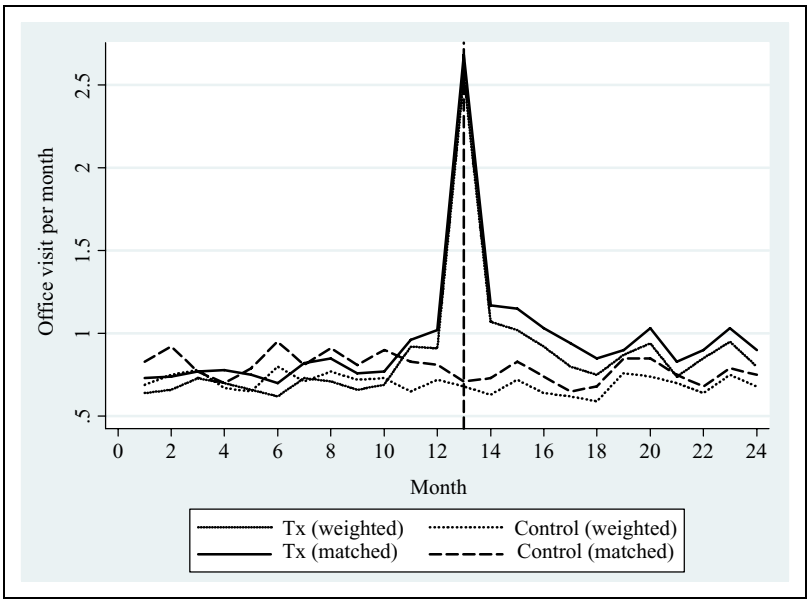

Figure 3. Adjusted monthly office visit rates for treatment and controls, using both propensity score-based weights and propensity score matched pairs. The vertical dashed line represents program initiation.

procedure must be conducted on the entire population at once (in effect, this holds all baseline characteristics constant between treated and potential controls). Therefore, this study design cannot be readily implemented until a sufficient number of treated units are enrolled in the intervention. If it is unfeasible to enroll the entire treatment cohort at one time, a viable alternative is to use a block enrollment strategy, where sufficiently large treatment cohorts are enrolled at intervals of perhaps 3 or 6 months. The propensity score estimation can then be performed for each cohort separately and the outcomes of each adjusted cohort can be tracked according to their own time line.

The propensity score concept is adaptable to many observational research endeavors, where a comparable control group is desired. For example, Haviland, Rosenbaum, and Nagin (2007) combined propensity score matching and group-based trajectory analysis to balance covariates within trajectory groups of the outcome variable, whereas Linden and Adams (in press) developed a propensity score-based weighting

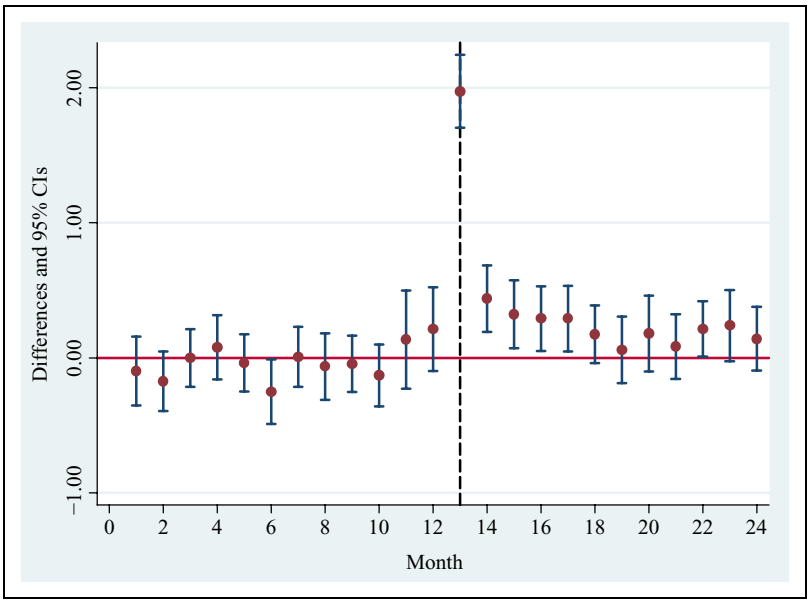

Figure 4. Estimates of program effect on monthly physician office visit rates. Values and 95% confidence intervals represent the treated group's monthly office visit rate relative to the propensity-scored matched controls (red line at zero). The vertical dashed line represents program initiation.

mechanism to be used with time-series data when only aggregated outcome data are available.

Limitations of the Proposed Framework

The most important assumption required when implementing any of propensity score-based adjustment methods described here is that there are no unmeasured confounders or biases remaining. This is a strong assumption, but this is the same assumption required to make a causal interpretation when estimating the effect of an intervention using any standard statistical methods. Unfortunately, there is no way to empirically validate this assumption from the data. That said, sensitivity analyses can be used to estimate the magnitude of hidden bias necessary to invalidate the study findings. The reader is referred to Rosenbaum (2002) for a comprehensive discussion on conducting sensitivity analyses in observational studies, and

Brumback, Hernán, Haneuse, and Robins (2004) for a more specific discussion on sensitivity analyses for longitudinal weighted models.

A general limitation of propensity scoring or, for that matter any other study design, pertains to the number of variables available to the evaluator for estimating propensity scores and assessing outcomes. The more variables available for use in estimating the propensity score, the more likely that a good fitting model can be developed while concomitantly reducing the amount of unmeasured confounders (Linden et al., 2005). Boosted logistic regression (Ridgeway, 1999) is worth considering as an alternative to the standard logistic model in estimating the propensity score. Regression boosting, commonly referred to as multiple additive regression trees (MART), is a general, automated, data-adaptive modeling algorithm that can estimate the nonlinear relationship between the outcome variable (in this case, treatment assignment) and a large number of covariates including multiple level interaction terms resulting in greater accuracy over standard linear models (McCaffrey, Ridgeway, & Morral, 2004). Using such algorithms induces balance on a greater number of measured covariates, making the groups similar in a more measured way.

One concern that many decision makers in the commercial health care industry have regarding the concept of matching is that all or nearly all treated units will find successful matches, whereas most of the nonparticipants in the population remain unmatched and thereby excluded from the analysis. By excluding the unmatched population from the analysis, the effect of nontreatment in the remaining population is not captured. Thus, we gain no insight as to how well the program chose its participants or if the program could have been effective among those individuals not explicitly targeted for the intervention (Linden & Adams, 2008). Stratification of the entire population into propensity score quintiles ameliorates this concern to a large degree by allowing us to review baseline and outcomes for participants, matched controls, and those individuals left completely unmatched. In fact, this last group can be viewed as a third cohort under study and compared to the other two groups (Linden & Adams, 2008). It should be noted, however, that the treatment effect estimated from stratification is the ATE, while matching estimates the ATT. Thus, decision maker should be aware that it is possible that different estimates will arise from these two methods.

Another limitation to propensity scoring techniques is that treated units must have scores different from zero (Linden & Adams, 2010a; Rosenbaum & Rubin, 1983). In effect, no treatment effect can be estimated for people who have no probability of receiving the treatment. A final limitation of

weighting adjustments is that they can perform poorly when the weights for a 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 (Linden & Adams, 2010a).

One drawback to the propensity score technique is the relatively low correlation that may accompany the use of covariates. An emerging idea is if the propensity score is not associated with outcomes, it becomes an instrument and should be used in instrumental variable analysis (Linden & Adams, 2006). Another possibility would be to use the propensity score within a regression-discontinuity design (Linden, Adams, & Roberts, 2006) as the pretest measure where a "cutoff" value would determine assignment. Although both of these ideas represent much more complicated modeling procedures, further thought should be given to their application in a prospective study format as proposed in this article.

Conclusion

This article describes a conceptual framework intended to emulate the randomization process using observational data, thereby allowing health care administrators to track initiatives in a prospective manner. Although these methods can never ensure the same level of validity as an RCT, they are considered robust alternatives when randomization is impractical. Following these techniques allows outcomes to be observed shortly after launching the initiative rather than waiting until study completion. The obvious benefit to stakeholders is that significant treatment effects can be observed as they occur, or alternatively, the initiative can be cancelled if treatment effects are not attained by a certain time point. As illustrated in the example provided, a significant treatment effect was realized immediately in the first month of the program but lasted for only 6 months. Thus decision makers would have a choice to either terminate the program or continue tracking the outcomes until a more comprehensive evaluation could be conducted at the end of the study. The concept proposed in this article is intended to offer a robust alternative to the inadequate strategies currently being used in the commercial health care industry where study findings may not be trusted, and thus decision makers remain uninformed as to whether an initiative is worth continuing or cancelled.

Declaration of Conflicting Interests

The author(s) declared no conflicts of interest with respect to the authorship and/or publication of this article.

Funding

The author(s) received no financial support for the research and/or authorship of this article.

References

Austin, P. C. (2008). A critical appraisal of propensity score matching in the medical literature from 1996 to 2003. *Statistics in Medicine*, *27*, 2037–2049.

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.

Brumback, B. A., Hernán, M. A., Haneuse, S. J. P. A., & Robins, J. M. (2004). Sensitivity analyses for unmeasured confounding assuming a marginal structural model for repeated measures. *Statistics in Medicine*, *23*, 749-767.

Chambers, J. M., Cleveland, W. S., Kleiner, B., & Tukey, P. A. (1983). *Graphical methods for data analysis.* Belmont, CA: Wadsworth.

Cochran, W. G. (1968). The effectiveness of adjustment by subclassification in removing bias in observational studies. *Biometrics*, *24*, 205-213.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.

Conover, W. J. (1999). *Practical nonparametric statistics* (3rd ed.). New York, NY: John Wiley.

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.

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

Freedman, D. (1999). From association to causation: Some remarks on the history of statistics. *Statistical Science*, *14*, 243-258.

Haviland, A., Rosenbaum, P. R., & Nagin, D. S. (2007). Combining propensity score matching and group-based trajectory analysis in an observational study. *Psychological Methods*, *12*, 247-267.

Heckman, J., Ichimura, J., & Todd, P. (1997). Matching as an econometric evaluation estimator: Evidence from evaluating a job training programme. *Review of Economic Studies*, *64*, 605-654.

Hochberg, Y., & Tamhane, A. C. (1987). *Multiple comparison procedures.* New York, NY: John Wiley.

Imai, K., King, G., & Stuart, E. A. (2008). Misunderstandings between experimentalists and observationalists about causal inference. *Journal of the Royal Statistical Society*, *171*, 481-502.

Imbens, G. (2004). Nonparametric estimation of average treatment effects under exogeneity: A survey. *Review of Economics and Statistics*, *86*, 4-30.

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

Linden, A., & Adams, J. L. (2008). Improving participant selection in disease management programs: Insights gained from propensity score stratification. *Journal of Evaluation in Clinical Practice*, *14*, 914-918.

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.

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.

Linden, A., & Adams, J. L. (in press). Applying a propensity-score based weighting model to interrupted time series data: Improving causal inference in program evaluation. *Journal of Evaluation in Clinical Practice*.

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*, 107-127.

Linden, A., Adams, J., & Roberts, N. (2006). Evaluating disease management program effectiveness: An introduction to the regression-discontinuity design. *Journal of Evaluation in Clinical Practice*, *12*, 124-131.

McCaffrey, D., Ridgeway, G., & Morral, A. (2004). Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychological Methods*, *9*, 403-425.

Nichols, A. (2008). Erratum and discussion of propensity-score reweighting. *Stata Journal*, *8*, 532-539.

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.

Ridgeway, G. (1999). The state of boosting. *Computing Science and Statistics*, *31*, 172-181.

Robins, J. M., Hernán, M. A., & Brumback, B. (2000). Marginal structural models and causal inference in epidemiology. *Epidemiology*, *11*, 550-560.

Rosenbaum, P. (2002). *Observational studies* (2nd ed.). New York, NY: Springer.

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

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.

Rosenbaum, P. R., & Rubin, D. B. (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *American Statistician*, *39*, 33-38.

Rubin, D. B. (1980). Bias reduction using Mahalanobis metric matching. *Biometrics*, *36*, 293-298.

Rubin, D. B. (2007). The design versus the analysis of observational studies for causal effects: Parallels with the design of randomized trials. *Statistics in Medicine*, *26*, 20-30.

Rubin, D. B., & Thomas, N. (1996). Matching using estimated propensity scores: Relating theory to practice. *Biometrics*, *52*, 249-264.