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

Journal of Evaluation in Clinical Practice ISSN 1365-2753

Combining the regression discontinuity design and propensity score-based weighting to improve causal inference in program evaluation

Ariel Linden DrPH[1,2] and John L. Adams PhD[3]

[1]President, Linden Consulting Group, Ann Arbor, MI, USA
[2]Adjunct Associate Professor, Department of Health Policy & Management, School of Public Health, University of Michigan, Ann Arbor, MI, USA
[3]Senior Statistician, RAND Corporation, Santa Monica, CA, USA

Keywords
covariate balance, inverse probability of treatment weight, observational studies, propensity score, regression discontinuity

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

Accepted for publication: 16 August 2011

doi:10.1111/j.1365-2753.2011.01768.x

Abstract

The regression discontinuity (RD) design is considered to be the closest to a randomized trial that can be applied in non-experimental settings. The design relies on a cut-off point on a continuous baseline variable to assign individuals to treatment. The individuals just to the right and left of the cut-off are assumed to be exchangeable – as in a randomized trial. Any observed discontinuity in the relationship between the assignment variable and outcome is therefore considered evidence of a treatment effect. In this paper, we describe key advances in the RD design over the past decade and illustrate their implementation using data from a health management intervention. We then introduce the propensity score-based weighting technique as a complement to the RD design to correct for imbalances in baseline characteristics between treated and non-treated groups that may bias RD results. We find that the weighting strategy outperforms standard regression covariate adjustment in the present data. One clear advantage of the weighting technique over regression covariate adjustment is that we can directly inspect the degree to which balance was achieved. Because of its relative simplicity and tremendous utility, the RD design (either alone or combined with propensity score weighting adjustment) should be considered as an alternative approach to evaluate health management program effectiveness when using observational data.

1. Introduction

Regression discontinuity (RD) represents one of the strongest quasi-experimental designs available in observational studies because it relies on a cut-off point on a continuous baseline variable to assign individuals to treatment. The individuals just to the right and left of the cut-off are assumed to be exchangeable – as in a randomized trial. However, use of the RD design has been somewhat limited, due in part to challenges to the key assumption that the treatment assignment variable alone ensures balance (comparability) of other baseline covariates. One solution is to apply propensity scoring techniques [1] to adjust for observed differences on baseline characteristics. There has been some debate about whether propensity scores can serve this role in the RD design. In this paper, we suggest that the propensity score approach is in fact compatible with the RD design and provide a detailed example using data from a health management intervention. By doing so, we hope to facilitate broader use of the RD design as a robust technique that can be used to evaluate an array of interventions when randomization is not feasible.

The paper is organized as follows. In Section 2, we describe the RD design in more detail and then argue for the role of the propensity score when covariate imbalance on baseline characteristics is encountered. In Section 3, we step back and discuss recent developments in the RD design to enable us, in Section 4, to describe how one particular propensity score technique – inverse probability of treatment weighting (IPTW) – can complement the RD design. In Section 5, we illustrate the implementation of the combined techniques using data from a recent study estimating the impact of a health management intervention on reducing health care costs in a chronically ill group of individuals. We close by discussing limitations and offering some concluding thoughts.

2. The RD design and a proposed role for propensity score-based weighting

Observational study designs typically strive to closely replicate a randomized controlled trial (RCT) by creating a control group that is essentially equivalent to the treatment group on observed baseline characteristics. This bolsters confidence in causal inferences about the effect of the treatment. The RD design, first described by Thistlethwaite & Campbell [2], is perhaps the closest to a randomized trial found in non-experimental settings. The concept relies on a cut-off point on a continuous baseline variable to assign individuals to treatment. The individuals just to the right and left of the cut-off are assumed to be exchangeable – as in a randomized trial. Thus, any observed discontinuity in the relationship between the assignment variable and outcome is considered evidence of a treatment effect.

The theoretical case made for the RD design's strong internal validity is based on the premise that if individuals do not have *precise* control over their assignment variable score, they cannot self-select into treatment. The inability to self-select implies that individuals close to either side of the cut-off should be comparable on all baseline characteristics (or stated differently, observed and unobserved characteristics should be continuous across the cut-off on the assignment variable). It therefore follows that in an RD design the evaluation of the outcome variable in the neighbourhood of the cut-off is 'as good as randomized' and should provide an unbiased estimate of the treatment effect [3].

The validity of the RD design rests on the assumption that the treatment assignment variable alone ensures balance (comparability) of other baseline covariates. However, there are likely to be situations in which this assumption fails, even in the presence of strict adherence to the cut-off (i.e. individuals do not manipulate their treatment assignment). It is therefore important for the evaluator to consider *ex ante* all other characteristics, other than the treatment, that may differ systematically between those individuals above and below the cut-off.

To illustrate, biomarkers are routinely used to diagnose chronic diseases with established thresholds serving as prompts to initiate or modify treatment. However, most chronic conditions also have *complications* – other diagnoses physiologically related to the primary condition [4]. These complications generally impact morbidity and mortality independent of the primary condition; therefore, it is not valid to assume that these complications are monotonically associated with the assignment variable (as we would expect in the context of the RD design). For example, assume that an RD design was employed to evaluate the effectiveness of a diabetes management program in reducing hospital days, where glycosylated haemoglobin [HbA1c] was the assignment variable and patients with values >9.0% assigned to treatment (9.0% is the threshold used in the 2011 Physician Quality Reporting System to indicate poor diabetes control) [5]. If there was a higher prevalence of cardiovascular disease (a primary complication of diabetes) in one of the groups, the estimated treatment effect would likely be biased against that group (as cardiovascular disease accounts for 24% of total hospital days attributable to diabetes) [6].

There are several techniques available to address the issue of balancing covariates in an RD design when the treatment assignment variable alone cannot ensure comparability of other baseline covariates. One approach is to apply inclusion/exclusion criteria as part of the data-processing step to ensure that the pool of study participants is relatively homogeneous on important characteristics prior to the statistical analysis. Using the diabetes example from above, the researcher could require that all individuals with an HbA1c in the neighbourhood of the cut-off (i.e. between 8.0 and 10.0%) have cardiovascular disease (or conversely, require that nobody does).

Alternatively, the researcher can leave the study population unaltered and apply covariate adjustment techniques as part of the evaluation process. The most commonly used adjustment method is to simply include baseline covariates in the outcome regression model. However, some have argued that the inclusion of baseline covariates in the regression model does nothing more than reduce the sampling variability [3,7], and in a broader sense, standard regression adjustment may elicit biased results, most notably when extrapolating between groups with completely non-overlapping data or in the presence of time-dependent confounders [8,9]. Another key limitation of covariate adjustment within a regression framework is that there is no way to validate whether the covariates have adequately adjusted for imbalances between groups.

Adjustment techniques based on the propensity score offer an attractive alternative to ensure covariate balance within the RD design framework. The propensity score, defined as the probability of assignment to the treatment group conditional on observed covariates [1], controls for baseline differences between treated and non-treated groups. Propensity scores are generally derived from a logistic regression equation that reduces each participant's set of covariates to a single score, ranging from 0 to 1.0. Referring back to the diabetes management example, the propensity score would be the probability of having an HbA1c value >9.0%, conditional on diabetes complications and all other observed characteristics. Conceptually, on average, any two individuals with the same propensity score will be balanced on all observed covariates, thereby reducing bias that could confound the estimated treatment effects. In contrast to regression-based adjustment, this method allows for covariate balance to be tested directly.

Imbens and Lemieux [7] contend that the propensity score approach may be at odds with the RD design because of the basic requirement that there be an overlap in all baseline covariates for which individuals are considered comparable. In the RD design, they argue, there is naturally no overlap in the assignment variable because of the strict cut-off, and therefore for all values of the assignment variable the probability of assignment is either 0 or 1, rather than in a range between 0 and 1.0 as required by the propensity score approach [7]. We, however, offer a somewhat different interpretation, and suggest that the propensity score approach is mutually compatible with the RD design. In keeping within the 'as good as randomized' framework of the RD design, we can consider the assignment variable in the neighbourhood of the cut-off as unassociated with the model (i.e. within the narrow range of the assignment variable we consider treatment assignment as equivalent to a coin toss). Thus, in estimating the propensity score model the researcher would exclude the assignment variable from the model (all other baseline covariates would be included as usual). If indeed individuals in the neighbourhood of the cut-off are exchangeable, then the resulting propensity score should provide the necessary overlap in covariates for ensuring that balance is achieved between groups.

In Section 4, we describe one particular propensity score technique, IPTW, and its mechanism for achieving balance between treated and non-treated groups on observed baseline characteristics. We then explain how the IPTW can complement the RD design for estimating the treatment effects of an intervention when there are differences between treatment and control groups on observed baseline covariates. In Section 5, we illustrate the implementation of the combined techniques using data from a recent health management intervention. Both of these sections require a deeper understanding of the techniques used when implementing the RD design. Therefore, in the next section we describe many of the key advances made in the RD design over the past decade.

3. Implementing the RD design: recent developments

The RD design historically involved estimation via standard parametric regression with the main treatment effect identified by a statistically significant *P*-value for the coefficient of the treatment variable [10–14]. A major drawback to this approach is that model specification is based on the entire range of the assignment variable and thus treatment effect estimates can be highly sensitive to observations far away from the cut-off. As a result, the researcher must devote substantial effort to ensuring that the functional form of the model is correctly specified along the entire continuum of observations (including fitting and testing higher level terms, interactions, etc.). One of the most significant advances made in the RD design over the past decade is in limiting the analysis to a range of scores in the neighbourhood around the cut-off. This increases the validity of the design as individuals within the neighbourhood should be most comparable. It also reduces the likelihood that researchers will incorrectly specify the functional form of the model. However, it increases the importance of determining the optimal size of the neighbourhood.

Estimation using non-parametric local linear regression

Recently, an alternative and more flexible estimation approach for RD has been suggested [15], which entails fitting two local linear regressions (LLRs), one on either side of the cut-off, and then predicting the value at the cut-off point. In general, LLR involves fitting a model linearly around a given point on the X variable (the assignment variable in the RD context) within a narrow range of the data surrounding that point (called a 'bandwidth' or what we call the neighbourhood) and applying a weighting scheme (called 'kernel weights') to down-weight the contributions of data points further away from the given X value. This process is performed across a series of X values in a grid, and then joined to obtain a smoothed curve [16].

When using LLR, the researcher is responsible for choosing both the kernel weight and bandwidth. There are several kernel weights to choose from; however, the triangular kernel is uniquely suited for RD because it is optimal for estimating LLRs at the boundary (which is the cut-off in RD) [16,17]. Selecting an optimal bandwidth is not as straightforward as choosing a kernel. Researchers can choose between 'rule-of-thumb' estimators [16], cross-validation procedures [7] or data-driven techniques [18]. In general, when the data are relatively linear, different bandwidths will likely elicit similar treatment estimates, and thus the choice of bandwidth selector is less important. However, when the data appear curvilinear, it is likely that treatment estimates will be very sensitive to the choice of bandwidth, increasing the importance of this choice. Lee and Lemieux [3] suggest exploring the sensitivity of the results to a range of bandwidths using various selectors, while McCrary [19] approaches the problem by estimating the treatment effect using bandwidths of half and twice the size of the basic bandwidth. Universally, commentators on the RD design advocate visual inspection of the prediction lines superimposed on a scatter plot of the actual data to informally assess the model fit at the various bandwidths.

Once the kernel weight and bandwidth have been chosen, the modelling process and derivation of treatment effects are straightforward. First, the LLR models are estimated separately to the left and right of the cut-off. The actual model entails regressing the outcome variable on the assignment variable, limited only to values on that respective side of the cut-off. Second, the predicted values for each model at the cut-off are stored. Third, the treatment effect estimate is obtained by subtracting the predicted value of the 'control' side of the cut-off from the predicted value of the 'treatment' side of the cut-off. Finally, this estimate is bootstrapped to derive non-parametric standard errors and confidence intervals (CIs).

Estimation using single model regression

While the LLR is a flexible alternative to a parametric modelling approach, the estimates derived from both procedures should be similar if the same bandwidth and kernel weights are applied. In the parametric model, this entails regressing the outcome (y) on the treatment variable (z), the assignment variable (x) and an interaction term between treatment and assignment (z * x). Once a bandwidth is chosen (via one of the approaches described above), a kernel weight is manually constructed and used in the regression model as a probability weight. For a triangle kernel, the weight equals the bandwidth minus the individual's assignment score. Thus, an assignment score closer to the cut-off gets weighted more heavily than assignment scores further away from the cut-off (with assignment scores beyond the bandwidth receiving a zero weight). The coefficient of the treatment parameter is the estimate of a treatment effect.

One consideration in choosing between a parametric versus non-parametric based approach is how standard errors are derived. In a parametric approach, the researcher can choose between robust analytic standard errors [20] to control for heteroscedasticity or by bootstrapping [21] the beta coefficient of the treatment parameter. In the non-parametric LLR approach, the researcher is limited to bootstrapping. One would not expect to find substantial differences in the standard errors derived from the two methods, but it would be beneficial to report both if they are dissimilar.

Robustness tests

Covariate balance

One of the first steps in any evaluation, whether randomized or observational, is to test whether treatment and control groups are comparable on baseline characteristics. Imbalances in covariates between groups can lead to systematic biases that may limit the validity of study findings. In the RCT, we assume that balance is naturally achieved in both observed and unobserved covariates. Because of selection bias, we cannot make this assumption in observational studies and must assess covariate balance based on observed characteristics. For RCTs and matching-type studies, there are several methods available to assess covariate balance including standardized differences [22], Kolmogorov–Smirnov

© 2012 Blackwell Publishing Ltd

equality of distributions test [23] or diagnostic plots such as quantile–quantile plots or box plots [24].

In an RD design, testing for covariate balance is conducted in a similar fashion to the method used for estimating treatment effects [3]. In the LLR technique, each baseline covariate is regressed on the assignment variable, limited only to values on that respective side of the cut-off, and the predicted values for each model at the cut-off are compared. Covariates are considered balanced when the value obtained by subtracting the predicted value of the 'control' side of the cut-off from the predicted value of the 'treatment' side of the cut-off is not statistically different from zero. As much balance as possible is desirable even if no statistically significant differences are found.

In using a parametric modelling approach, we regress the given baseline covariate on the treatment variable (z), the assignment variable (x) and an interaction term between treatment and assignment (z * x). The same bandwidth and kernel weights used in the outcome model are applied here. However, in contrast to the outcome model, the coefficient of the treatment parameter represents the estimate of covariate balance and is indicated by a non-statistically significant coefficient (and/or CIs that cross zero).

Manipulation of treatment assignment

In contrast to an RCT where treatment assignment is unknown to both administrators and participants prior to enrolment, in a study using an RD design there is the possibility that individuals could manipulate their treatment assignment. To do so, individuals would first have to know where the cut-off score is set, and then would need the ability to manipulate their own assignment score. To determine the likelihood that such manipulation occurred, McCrary [19] suggests testing the continuity in the density of the assignment variable at the cut-off. Returning to our diabetes management program example, assume that individuals with diabetes knew that the cut-off value for HbA1c was 9.0% and they could manipulate their blood glucose levels to get above that level. When reviewing the density of HbA1c scores, we would expect to see relatively few individuals with values just under 9.0% and relatively many individuals with values just over 9.0%. While a discontinuity in the density at the cut-off does not necessarily imply that such manipulation occurred, it does draw attention to the possibility and would warrant further investigation.

Testing for discontinuities away from the true cut-off

While a discontinuity at the cut-off may represent a true treatment effect, one would feel less sanguine about this result if discontinuities were also found elsewhere along the continuum of assignment values, especially at points where no effect is anticipated. Different approaches have been proposed to test for discontinuities away from the cut-off. Imbens and Lemieux [7] suggest dividing each subsample (to the left and right of the cut-off) at their respective median and testing for a discontinuity at the median. The median is a good choice of cut-off to maximize power to detect a significant jump (as the subsample will be evenly split on both sides). Nichols [25] suggests randomly choosing 100 placebo cut-off points from the range of the assignment variable and testing for a discontinuity. Using this approach, the underlying assumptions of the RD design may be considered violated if substantially more than 5% of these cases show a statistically significant discontinuity.

4. Propensity score-based weighting and RD

In this section, we return to our discussion of the propensity score as an approach to address covariate imbalance in the RD design. There are a variety of ways in which the propensity score can be used as the basis for deriving estimated treatment effects in observational studies. One method is to match treated and non-treated individuals on the propensity score and then conduct statistical analyses in the usual manner on the matched pairs alone. Another method is to construct a weight based on the conditional probability of an individual receiving his/her own treatment (referred to as the 'inverse probability of treatment weight' or IPTW) [9,26,27]. More specifically, participants receive a weight equal to the inverse of the estimated propensity score (1/propensity score), and non-participants receive a weight equal to the inverse of 1 minus the estimated propensity score (1/1–propensity score). This IPTW estimator sets the distribution of covariates to be equal to that of the population and thus estimates the average treatment effect (ATE). Other weighting schemes can be used to estimate different treatment effects, such as average treatment effect on the treated or average treatment effect on controls [28].

As a result of removing any existing association between baseline covariates and treatment, the IPTW creates a study population 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 baseline characteristics [29]; and (2) it offers 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) [9]. In essence, the IPTW weights the analysis so it looks as much as possible like a RCT.

The IPTW approach is a natural complement to the RD design. First, the IPTW is intended to provide an estimate of the ATE in the population for which treatment is appropriate [30]. This is perfectly aligned with the objective of the RD design which specifically estimates the ATE in the neighbourhood of the cut-off. Second, the weights can be easily added to any of the existing RD modelling strategies with little or no modification. For the single parametric modelling approach, this simply involves multiplying the IPTW weight by the kernel weight and using this new weight in the regression. In LLR models that generate and apply the kernel weights automatically, the IPTW weight is used without modification directly in the modelling process. The degree to which covariate balance is realized can then be tested using the method described in covariate balance.

5. Example: a health management program

In this section, we illustrate how the propensity score-based weighting technique can be combined with the RD design in the context of a health management program evaluation.

Table 1 The change in costs (program period – baseline) using an optimal bandwidth (100.09), 50% bandwidth and 200% bandwidth (*N* treated = 79, non-treated = 270)

	Unadjusted			Weighted		
Variable	Estimate	Low 95%	High 95%	Estimate	Low 95%	High 95%
Difference in cost (optimal)	3872.52	–3971.20	11 716.23	3619.19	–3957.21	11 195.60
Difference in cost (50%)	1986.38	–7627.24	11 600.00	3895.28	–7282.58	15 073.14
Difference in cost (200%)	1031.64	–5466.45	7 529.73	2917.06	–4061.27	9 895.38

Outcome estimate using regression with covariates = 4405.02 (95% confidence interval = –2655.69, 11 465.73).

Setting

Our data come from a primary care-based medical home pilot program that invited patients to enrol if they had a chronic illness or were predicted to have high costs in the following year. The goal of the pilot was to lower health care costs for program participants by providing intensified primary care that was intended to reduce unnecessary emergency department visits and hospitalizations.

Risk score

A baseline 'risk score' was calculated for all potential program participants, which indicated the expected relative cost risk of an individual compared with the population average. The risk score values ranged from 2 to 798 in the overall study population ($n = 2002$). To demonstrate the RD design, in the present study the risk score is used as the assignment variable. We chose the median risk score of the treatment group (198.5) as the cut-off, dropping all program participants with a score below the cut-off and all non-participants with scores above this cut-off. This produced a total sample size of 1664 individuals (184 treated and 1480 non-treated).

Propensity score estimation and IPTW weights

A propensity score was estimated using logistic regression to predict program participation status conditional on baseline demographic characteristics (age and gender); utilization of health services (primary care visits, other outpatient visits, laboratory tests, radiology tests, prescriptions filled, hospitalizations, emergency department visits); and total medical costs (the amount paid for all the health services). Each individual then received an IPTW based on their actual treatment status and estimated propensity score. Participants received a weight equal to the inverse of the estimated propensity score (1/propensity score), and non-participants received a weight equal to the inverse of 1 minus the estimated propensity score (1/1–propensity score).

Outcome measure

The outcome of interest was the change in total medical costs from the 12 months prior to the program (baseline period) to the 12 months after program initiation (program period) in treatment versus control groups. This approach to measuring the outcome is referred to as a differences-in-differences (DID) estimator. A positive value for the DID estimate indicates that the program participant group had an increase in costs greater than the non-participant group, and a negative value indicates that the program participant group had a decrease in costs greater than non-participants. The DID strategy ensures that any variables that remain constant over time (but are unobserved) will not bias the estimated effect [31].

Results

Outcomes

Table 1 presents the results of the unadjusted and weighted RD regression analysis for the outcome variable – the change in total costs. Data were analysed using the single model parametric approach described earlier. Additionally, to test the sensitivity to the choice of bandwidth, we followed the approach suggested by McCrary [19] by estimating the treatment effect using bandwidths of half and twice the size of the chosen bandwidth. An optimal bandwidth of 100.09 was determined using the data-driven technique proposed by Imbens and Kalyanaraman [18] with a triangle kernel weight.

As shown, the unadjusted marginal difference in pre-to-post costs for the treatment group was $3872 *higher* than the non-treated group. In other words, the treated group's costs increased relative to the non-treated group. However, this estimate is not statistically significant as indicated by the 95% CIs crossing zero.

The weighted estimate and CIs (at the optimal bandwidth level) were very similar to the unadjusted model results. For comparison, we also estimated the difference in costs using a regression model employing all the covariates originally used to estimate the propensity score (in lieu of using the propensity score-based weight). This approach provided a point estimate of $4405.02 which was $785 higher than the weighted model. Additionally, the CI (95% CI: −2655.69 to 11 465.73) was somewhat wider than that of the weighted model. Taken together, these findings increase our confidence that the weighting technique is a reasonable approach to adjusting for baseline characteristics without substantially altering the magnitude of the outcome.

Also of note is that the weighted estimates at the various bandwidths were more consistent than those for the unadjusted models. This suggests that the weighted adjustment may be less sensitive to bandwidth selection, though future study using simulations of various sample size and bandwidths is warranted.

Figure 1 provides a visual display of the data and unadjusted estimates. The prediction lines for left and right sides of the cut-off were produced using LLR with the optimal bandwidth of 100.09 and a triangle kernel weight. These are superimposed on the x-y scatter of the actual data points. One would be hard pressed to discern any clear discontinuity within the local neighbourhood of

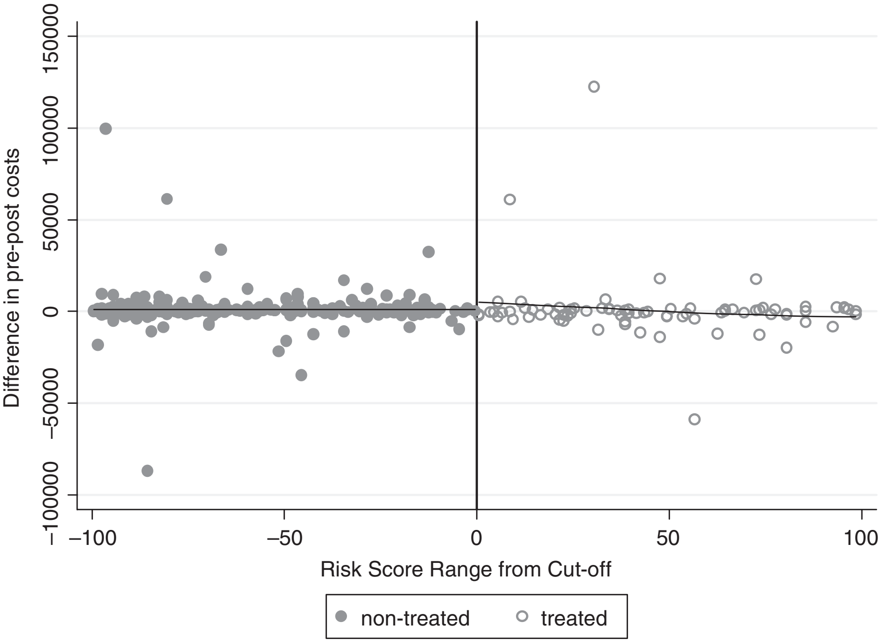

Figure 1 Difference in pre-to-post costs plotted against a risk score range within a bandwidth of 100.09 points of the cut-off. Values to the left of zero are for the non-participants and values to the right are for program participants.

the cut-off. This illustration lends support to the unadjusted estimates reported in Table 1. The weighted data appeared nearly identical to the unadjusted data displayed in Fig. 1 (not shown).

Following the suggested approach of Imbens and Lemieux [7], we also tested for discontinuities away from the actual cut-off by dividing each subsample (to the left and right of the cut-off) at their respective median values of the risk score (39 for the non-participant group and 313.5 for the program participant group). No statistically significant discontinuities were found at either point (data not shown).

Covariate balance

As described in covariate balance, the general approach to testing for covariate balance in the RD design involves regressing a given baseline covariate on the assignment variable. Covariates are considered balanced when the estimate is not statistically different from zero (or in the case of CIs, the values cross zero).

Table 2 provides unadjusted and weighted estimates for each of the baseline covariates, estimated using the single regression model with an optimal bandwidth of 100.09 and triangle kernel. As shown, in the unadjusted approach there were several baseline covariates with statistically significant imbalances (age, female, primary care visits, laboratory tests and prescription refills). However, once the weighting approach was applied, balance was achieved across all covariates. These estimates are similar to those found in the original evaluation data when using a matching approach [32].

Testing for manipulation of treatment assignment

Figure 2 illustrates the density of the assignment variable (risk score) in the neighbourhood of the cut-off together with prediction lines and 95% CIs, using the approach suggested by McCrary [19]. As shown, there is no discontinuity at the cut-off. This is not surprising as we set the cut-off at a rather arbitrary point in the data specifically to demonstrate the RD design. Thus, there is no reason to assume that individuals would (or could) manipulate their risk score to get into the treatment group (or vice versa).

6. Discussion

In this paper, we have described and implemented recent advances in the RD design as well as demonstrated how the propensity score-based weighting technique complements the design. As an additional enhancement to the overall evaluation strategy, a DID estimator was used to control for time-constant unobserved characteristics which may be correlated with the covariates in the model. In these data, the weighting mechanism balanced the observed baseline covariates, while in the unadjusted (conventional) model the covariates remained imbalanced. Moreover, the weighting mechanism outperformed the standard regression covariate adjustment approach, which was evidenced by less variability and a treatment effect estimate closer to the unadjusted (conventional) RD treatment effect estimates. The results found here were similar to the outcomes reported in Linden [32] that used the same dataset with a propensity score matching approach.

Generalizability of results

Given that the present data were intended solely for illustrating the proposed weighting approach, it is important to verify whether these results can be replicated in other data. As a simple additional experiment, we generated an artificial data set with 1000 observations, an outcome variable with a large treatment effect and a covariate with a large imbalance. The results of the simulation supported those of the example described in this paper. The mean

Table 2 Raw and weighted baseline (12 months prior to program participation) covariates (optimal bandwidth = 100.09) (*N* treated = 79, non-treated = 270)

	Unadjusted			Weighted		
Variable	Estimate	Low 95%	High 95%	Estimate	Low 95%	High 95%
Demographic characteristics						
Age	6.96	2.10	11.82	3.01	–4.11	10.14
Female	0.29	0.03	0.55	–0.04	–0.42	0.34
Utilization and cost						
Primary care visits	3.32	0.41	6.23	–1.71	–7.27	3.85
Other outpatient visits	0.17	–0.58	0.91	–0.72	–3.02	1.58
Laboratory tests	2.54	0.42	4.65	–0.11	–2.55	2.33
Radiology tests	0.80	–0.95	2.55	–0.43	–3.34	2.47
Prescriptions filled	18.97	2.84	35.10	–8.44	–35.29	18.40
Hospitalizations	0.10	–0.09	0.29	0.06	–0.12	0.25
Emergency department visits	0.03	–0.46	0.51	–0.23	–0.95	0.49
Total costs	3356.03	–808.81	7520.88	–193.93	–3771.35	3383.48

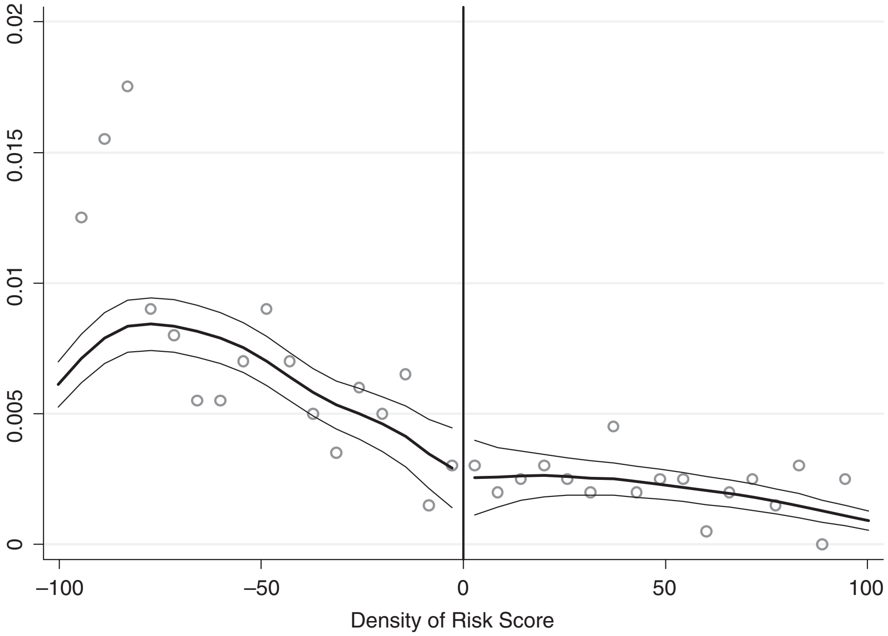

Figure 2 Density of the risk score (assignment variable). Values to the left of zero are for the non-participants and values to the right are for program participants.

treatment effect of the weighted model differed by only 1.0% from the conventional RD model while achieving balance on the covariate and exhibiting nearly identical CIs. On the other hand, the RD model estimated with the covariate was 4.0% different than the conventional RD model result and the CIs were slightly wider (data available from the first author). While these results are promising, future work should examine the robustness of the approach under many different scenarios (i.e. varying sample size, effect size, variability, magnitude of the covariate bias, etc.).

Limitations

Many of the problems arising in the RD design have been addressed earlier in the paper, such as model misspecification, sensitivity to the choice of bandwidth, potential manipulation of the assignment variable, etc.

Another potential limitation of the design may occur in settings using a temporal assignment variable (such as age or calendar date) with a cut-off after which everyone receives the treatment. While in such situations individuals cannot directly manipulate their assignment variable (unless of course they lie about their age to receive the treatment ahead of schedule), they can behave differently prior to crossing the cut-off in anticipation of receiving the treatment. The biases inherent in this set-up are similar to those in any simple pre-post study [33]. For example, Card *et al*. [34] found that individuals with little or no insurance coverage prior to age 65 tended to increase their number of routine doctor visits after enrolling in Medicare (reflective of moral hazard), while individuals with comprehensive insurance prior to Medicare did not exhibit a discontinuity in the number of routine doctor visits across the age 65 cut-off. However, by comparing outcomes across various socio-economic groups, Card

[34] avoided many of the threats to validity common in single group pre-post studies.

A limitation related to the weighting adjustment is that it 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 [27].

The 'fuzzy' design

In this paper, we described the RD design in the context of situations where there is strict adherence to the cut-off (generally referred to as the 'sharp' RD design). That is, all individuals on one side of the cut-off receive no treatment and all those on the other side of the cut-off receive the treatment (from a statistical standpoint this means that the probability of receiving treatment changes sharply from 0 to 1 at the cut-off). While beyond the scope of this paper, there are situations where some individuals on either side of the cut-off may receive the alternate treatment assignment. Referring back to our diabetes management example from earlier in the paper, some of those individuals meeting the enrolment criteria (HbA1c >9.0%) will refuse to participate. On the other hand, some individuals may be allowed to enrol in the program even though their HbA1c levels are lower than the cut-off criteria (perhaps because they were referred by their health care provider as exceptional cases). Trochim [10] labelled this condition the 'fuzzy' RD design. For a comprehensive discussion on the fuzzy design, the reader should refer to References [3], [7] and [15].

7. Conclusion

In this paper, we have described many of the key advances made in the RD design over the past decade and illustrated their implementation using data from a health management intervention. Additionally, we have presented the propensity score-based weighting technique as a complement to the RD design to correct for imbalances in baseline characteristics between treated and non-treated groups. Our results suggest that the weighting strategy outperforms standard regression covariate adjustment; however, this should be confirmed using other datasets and simulations. One clear advantage of the weighting technique over simply including these baseline covariates in the outcome regression model is that we can directly inspect the degree to which balance was achieved. Because of its relative simplicity and tremendous utility, the regression discontinuity design (either alone or combined with propensity score weighting adjustment) should be considered as an alternative procedure for use with observational data to evaluate health management program effectiveness.

8. References

1. Rosenbaum, P. R. & Rubin, D. B. (1983) The central role of propensity score in observational studies for causal effects. *Biometrika*, 70, 41–55.
2. Thistlethwaite, D. & Campbell, D. (1960) Regression-discontinuity analysis: an alternative to the ex post facto experiment. *Journal of Educational Psychology*, 51, 309–317.
3. Lee, D. S. & Lemieux, T. (2010) Regression discontinuity designs in econometrics. *Journal of Economic Literature*, 48, 281–355.
4. Linden, A., Biuso, T. J., Gopal, A., Barker, A. F., Cigarroa, J., Haranath, S. P., Rinkevich, D. & Stajduhar, K. (2007) Consensus development and application of ICD-9 codes for defining chronic illnesses and their complications. *Disease Management and Health Outcomes*, 15 (5), 315–322.
5. Centers for Medicare and Medicaid Services (2011) Physician Quality Reporting System (Physician Quality Reporting): Measures Groups Specifications Manual. Available at: https://www.cms.gov/pqrs/downloads/2011_PhysQualRptg_MeasuresGroups_SpecificationsManual_033111.pdf?agree=yes&next=Accept (last accessed 29 June 2011).
6. American Diabetes Association (2003) Economic costs of diabetes in the US in 2002. *Diabetes Care*, 26 (3), 917–932.
7. Imbens, G. W. & Lemieux, T. (2008) Regression discontinuity designs: a guide to practice. *Journal of Econometrics*, 142 (2), 615–635.
8. Freedman, D. (1999) From association to causation: some remarks on the history of statistics. *Statistics in Science*, 14, 243–258.
9. Robins, J. M., Hernán, M. A. & Brumback, B. (2000) Marginal structural models and causal inference in epidemiology. *Epidemiology*, 11, 550–560.
10. Trochim, W. M. K. (1984) Research Design for Program Evaluation: The Regression-Discontinuity Approach. Beverly Hills, CA: Sage Publications.
11. Trochim, W. M. K. (1990) The regression discontinuity design. In Research Methodology: Strengthening Causal Interpretations of Non-experimental Data. AHCPR Conference Proceedings, *PHS 90-3545* (eds L. B. Sechrest, E. Perrin & J. Bunker), pp. 119–139. Rockville, MD: Agency for Health Care Policy and Research.
12. Trochim, W. M. K., Cappelleri, J. C. & Reichardt, C. S. (1991) Random measurement error does not bias the treatment effect estimate in the regression-discontinuity design: II. When an interaction effect is present. *Evaluation Review*, 15, 571–604.
13. Shadish, S. R., Cook, T. D. & Campbell, D. T. (2002) Experimental and Quasi-Experimental Designs for Generalized Causal Inference. Boston, MA: Houghton Mifflin.
14. 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 (2), 124–131.
15. Hahn, J., Todd, P. & van der Klaauw, W. (2001) Identification and estimation of treatment effects with a regression discontinuity design. *Econometrica*, 69, 201–209.
16. Fan, J. & Gijbels, I. (1996) Local Polynomial Modelling and Its Applications. London; New York and Melbourne: Chapman and Hall.
17. Cheng, M.-Y., Fan, J. & Marron, J. S. (1997) On automatic boundary corrections. *The Annals of Statistics*, 25 (4), 1691–1708.
18. Imbens, G. W. & Kalyanaraman, K. (2009) Optimal Bandwidth Choice for the Regression Discontinuity Estimator. National Bureau of Economic Research Working Paper 14726.
19. McCrary, J. (2008) Manipulation of the running variable in the regression discontinuity design: a density test. *Journal of Econometrics*, 142 (2), 698–714.
20. White, H. A. (1980) A heteroscedasticity-consistent covariance matrix estimator and a direct test of heteroscedasticity. *Econometrica*, 48, 817–838.
21. 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.
22. Flury, B. K. & Reidwyl, H. (1986) Standard distance in univariate and multivariate analysis. *The American Statistician*, 40, 249–251.
23. Conover, W. J. (1999) Practical Nonparametric Statistics, 3rd edn. New York: Wiley.

24. Chambers, J. M., Cleveland, W. S., Kleiner, B. & Tukey, P. A. (1983) Graphical Methods for Data Analysis. Belmont, CA: Wadsworth.
25. Nichols, A. (2007) Causal inference with observational data. *Stata Journal*, 7 (4), 507–541.
26. Robins, J. M. (1998) Marginal structural models. 1997 Proceedings of the Section on Bayesian Statistical Science, pp. 1–10. Alexandria, VA: American Statistical Association.
27. 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.
28. Nichols, A. (2008) Erratum and discussion of propensity-score reweighting. *Stata Journal*, 8 (4), 532–539.
29. Rosenbaum, P. R. (1987) Model-based direct adjustment. *Journal American Statistical Association*, 82, 387–394.
30. Imai, K., King, G. & Stuart, E. A. (2008) Misunderstandings between experimentalists and observationalists about causal inference. *Journal of the Royal Statistical Society*, 171 (Part 2), 481–502.
31. Buckley, J. & Shang, Y. (2003) Estimating policy and program effects with observational data: the 'differences-in-differences' estimator. *Practical Assessment, Research & Evaluation*, 8 (24). Available at: http://PAREonline.net/getvn.asp?v=8&n=24 (last accessed 14 June 2011).
32. Linden, A. (2011) Identifying spin in health management evaluations. *Journal of Evaluation in Clinical Practice*, 17 (6), 1223–1230.
33. Linden, A., Adams, J. & Roberts, N. (2003) An assessment of the total population approach for evaluating disease management program effectiveness. *Disease Management*, 6 (2), 93–102.
34. Card, D., Dobkin, C. & Maestas, N. (2008) The impact of nearly universal insurance coverage on health care utilization: evidence from Medicare. *American Economic Review*, 98 (5), 2242–2258.