Received: 26 May 2022 | Revised: 16 May 2023 | Accepted: 14 July 2023

DOI: 10.1002/sim.9861

TUTORIAL IN BIOSTATISTICS

Statistics in Medicine WILEY

A guide to regression discontinuity designs in medical applications

Matias D. Cattaneo[1] | **Luke Keele**[2] | **Rocío Titiunik**[3]

[1]Dept. of Operations Research and Financial Engineering, Princeton University, Princeton, New Jersey, USA

[2]Dept. of Surgery, University of Pennsylvania, Philadelphia, Pennsylvania, USA

[3]Dept. of Politics, Princeton University, Princeton, New Jersey, USA

Correspondence
Luke Keele, Dept. of Surgery, University of Pennsylvania, 3400 Spruce St., Philadelphia, PA 19104, USA.
Email: luke.keele@gmail.com

Funding information
Foundation for the National Institutes of Health, Grant/Award Number: R01 GM072611-16; National Science Foundation, Grant/Award Numbers: SES-2019432, SES-2241575

We present a practical guide for the analysis of regression discontinuity (RD) designs in biomedical contexts. We begin by introducing key concepts, assumptions, and estimands within both the continuity-based framework and the local randomization framework. We then discuss modern estimation and inference methods within both frameworks, including approaches for bandwidth or local neighborhood selection, optimal treatment effect point estimation, and robust bias-corrected inference methods for uncertainty quantification. We also overview empirical falsification tests that can be used to support key assumptions. Our discussion focuses on two particular features that are relevant in biomedical research: (i) fuzzy RD designs, which often arise when therapeutic treatments are based on clinical guidelines, but patients with scores near the cutoff are treated contrary to the assignment rule; and (ii) RD designs with discrete scores, which are ubiquitous in biomedical applications. We illustrate our discussion with three empirical applications: the effect CD4 guidelines for anti-retroviral therapy on retention of HIV patients in South Africa, the effect of genetic guidelines for chemotherapy on breast cancer recurrence in the United States, and the effects of age-based patient cost-sharing on healthcare utilization in Taiwan. Complete replication materials employing publicly available data and statistical software in `Python`, `R` and `Stata` are provided, offering researchers all necessary tools to conduct an RD analysis.

KEYWORDS
causal inference, natural experiments, regression discontinuity

1 | INTRODUCTION

Drawing causal inferences from quantitative data is a fundamental goal in epidemiology, comparative effectiveness, health services, and outcomes research.[1-3] It is now well understood that while randomized controlled trials are the gold standard for learning about treatment effects, reliance on observational studies is unavoidable—there are simply too many contexts where randomization is infeasible or unethical. When randomization is not possible, evidence from natural experiments is often viewed as the next best alternative for causal inference and program evaluation.[3-6] Some scholars have advocated for greater use of the regression discontinuity (RD) design in biomedical contexts,[7-10] which can be viewed as a prime example of a natural experiment.[11] As a result, RD designs have become more common in biomedical research: a recent review identified over 325 studies based on RD designs in medical studies alone.[12]

Abbreviation: RD, regression discontinuity.

The popularity of the RD design stems from its high internal validity. Causal inferences from RD designs are often more credible and robust than those from other nonexperimental impact evaluation strategies such as selection-on-observables, difference-in-difference, or instrumental variable (IV) designs. The feature that contributes to the superior credibility of the RD design is the existence of an objective and verifiable treatment assignment rule that offers a design-based way to validate some of its key assumptions. In the canonical RD design, each unit i receives a score X_i, and a treatment is assigned according to the rule $T_i = \mathbb{1}(X_i \geq c)$, where c is a fixed known cutoff and $\mathbb{1}(\cdot)$ the indicator function, so that all units with score above the cutoff are assigned to the active treatment condition ($T_i = 1$) and all units with scores below the cutoff are assigned to the control condition ($T_i = 0$). The score X_i can be continuous (each unit has a unique score value) or discrete (multiple units share the same score value). In the so-called *sharp* RD design, all units comply perfectly with the treatment condition they are assigned: no units below the cutoff receive the treatment and no units above the cutoff refuse the treatment. In the more general case, referred to as the *fuzzy* RD design, the treatment assignment rule induces many units to take the treatment, but compliance with the assignment is imperfect. Each variant of the RD design requires conceptually and methodologically different approaches for analysis.

The RD design was first introduced by Thistlethwaite and Campbell[13] in education research to study the effect of receiving a certificate of merit based on test scores. In biomedical contexts, RD designs naturally arise from treatment guidelines based on diagnostic test results. For instance, a specific treatment is recommended when test results exceed a known cutoff—for example, start blood pressure medication when systolic blood pressure is above 130 mmHg. The key idea behind the RD design is that units just above and just below the cutoff should be comparable in terms of all unobservable and observable characteristics not affected by the treatment, which in turn implies that these units' differences in outcomes can be understood as the result of differences in treatment status rather than the result of differences in observable or unobservable characteristics. For example, assuming that patients do not have precise control over their blood pressure measurement, patients whose systolic pressure is 130 mmHg should be similar to patients whose systolic pressure is 129 mmHg: in a small neighborhood around the 130 cutoff, patients' particular measures will be governed by random chance (variable device accuracy, inadequate arm support, elevated anxiety, etc.) more than by patients' underlying health risks or other confounding factors affecting the outcome of interest.

We provide a systematic overview of the state-of-the-art statistical methodologies to analyze and interpret RD designs employing the two most used methodological frameworks: the continuity-based framework and the local randomization framework. Our discussion covers key assumptions, estimation methods, inference procedures, and diagnostic tests complementing and expanding on early introductory articles for the biomedical sciences,[7,8,10] which do not discuss the most recent RD methods that are now widely used in the statistical, social, and behavioral sciences.[14-16] Furthermore, we discuss two complications that frequently arise in biomedical research that have not been addressed by prior biomedical reviews: imperfect treatment compliance and noncontinuous score variables.

The manuscript is organized as follows. Sections 2 and 3 focus on RD methodology when the score is (approximately) continuous; for illustration, we reanalyze and expand a recent study that used the RD design to estimate the effect of immediate vs deferred anti-retroviral therapy (ART) on retention in care.[17] In this application, the score (CD4 count) takes on many distinct values and thus may be analyzed using RD methods suitable for continuous scores. Section 4 then discusses RD designs with a discrete score, that is, settings where the score takes on at most a few distinct values. We overview how the methods for RD designs with a (approximately) continuous score can be modified and extended, illustrating our discussion with two additional empirical examples. One example looks at genetic guidelines for chemotherapy and serves mostly as a cautionary tale because the key RD assumptions are not supported empirically. The other example studies patient cost-sharing and healthcare utilization and showcases how RD methods with discrete scores can be deployed successfully. Finally, Section 5 summarizes key takeaways for practice and concludes. The online materials include the three data sets as well as computer code to replicate all our analyses. Replication codes are available in `Python`, `R`, and `Stata`, and can be found at https://rdpackages.github.io/. The supplementary materials also include code to demonstrate a basic RD analysis.

2 | SETUP AND TREATMENT EFFECTS

In the canonical RD design, there are $i = 1, 2, \ldots, n$ units of analysis, each unit receives a *score* X_i (also known as *running variable*, *forcing variable*, or *index*), and a binary treatment is assigned based on whether this score exceeds or not a known cutoff c: units whose score is above the cutoff are assigned to the treatment condition, and units whose score is below the cutoff are assigned to the control condition. Thus, the probability of treatment assignment

as a function of the score changes discontinuously at the cutoff: all units above the cutoff are assigned to the treatment condition with probability one, while all units below the cutoff are assigned to the control condition with probability one. These three elements—score, cutoff, and treatment—are the key components of all RD designs. Crucially, the RD treatment assignment rule is known, at least to the researcher, and hence empirically verifiable. This distinctive feature contributes to the RD design's superior credibility when compared to other nonexperimental methods.

2.1 | Empirical example: ART and retention in care

We revisit the recent study by Bor et al,[17] who used a RD design to estimate the effect of immediate (vs deferred) anti-retroviral therapy (ART) on retention in care. The authors analyzed the Hlabisa HIV Treatment and Care Programme in South Africa, conducted by the Africa Health Research Institute and the South African Department of Health. This program collected data on all patients receiving HIV care and treatment services at government facilities (17 clinics and 1 hospital) between 12 August 2011 and 31 December 2012.[18,19] Patients were eligible for ART if their CD4 count was less than 350 cells/μl, and they had a WHO stage III/IV condition. Patients did an initial blood draw for a CD4 count, and were instructed to return to the clinic in one week to receive their result. ART-eligible patients were enrolled in several weeks of counseling and were then initiated on ART.

The investigators compared differences in retention between patients with CD4 counts (X_i) just above vs just below the 350-cells/μl threshold (c). The cohort included $n = 11\ 306$ patients and the data contained information on several predetermined covariates, including sex, age, date of testing, and testing location. This is a prototypical biomedical RD example, where the units of analysis are patients and the score (CD4 count) is the result of a diagnostic test. The cutoff is 350 cells/μl and the treatment is the immediate initiation of ART. The outcome of interest is a binary variable with value 1 if there was any evidence of any routine clinic visits, lab result (CD4 or viral load), or date of ART initiation 6 to 18 months after a patient's first CD4 count, regardless of receipt of ART. The RD design is fuzzy because not all patients with a score of less than 350 initiated ART (see Figure 3 in the next section). Henceforth, we refer to this example as the *ART* application.

Note that the treatment is assigned when the CD4 count is below (rather than above) the 350 cutoff. However, the score and cutoff can always be redefined so that treatment assignment occurs when $X_i \geq c$ (ie, multiplying both variables by -1, a relabeling of the data with no substantive effects in the analysis). Alternatively, the analysis can proceed given the original setup, and treatment effects are simply understood with a change of sign.

2.2 | Graphical illustration of RD design

An important first step in RD analysis is a graphical illustration of the design. Figure 1 showcases two basic plots. When properly executed, a graphical RD analysis adds transparency and credibility by displaying the observations used for estimation and inference, both globally (over the entire support of the score) and locally (near the cutoff determining treatment assignment). RD plots can also highlight other features of the design such as the coarseness of the score and outcome variables, the variability of the data, and the potential curvature of the underlying regression functions.[20] Despite their visual usefulness, RD plots should not be used as the main tool for the analysis, as they can often be misleading;[21] their main role should be as supplementary to the formal statistical analyses that we discuss in Section 3.

Figure 1A depicts a histogram of the score variable X_i, which captures the relative frequency of different observed values by first binning its support. In general, the score can be continuously or discretely distributed. When the score is continuous each unit has a unique score value, while when the score is discrete several units share the same score value and thus the score exhibits "mass points." In the ART application, X_i takes on $K = 1229$ distinct values in a sample of 11 306 total observations, so several observations share the same score value ($K < n = 11\ 306$). Given the sizable number of distinct values, we treat this application as having an approximately continuous score and discuss RD methods for that context. This is a common approach in practice when the discrete score has "many" mass points,[14,16] and was the approach adopted by Bor et al.[17] In Section 4, we discuss RD methods appropriate in cases when the score is discrete with possibly only a "few" distinct values. We also use Figure 1A as the starting point for validation of the RD design via discontinuity-in-density testing[22] in Section 3.

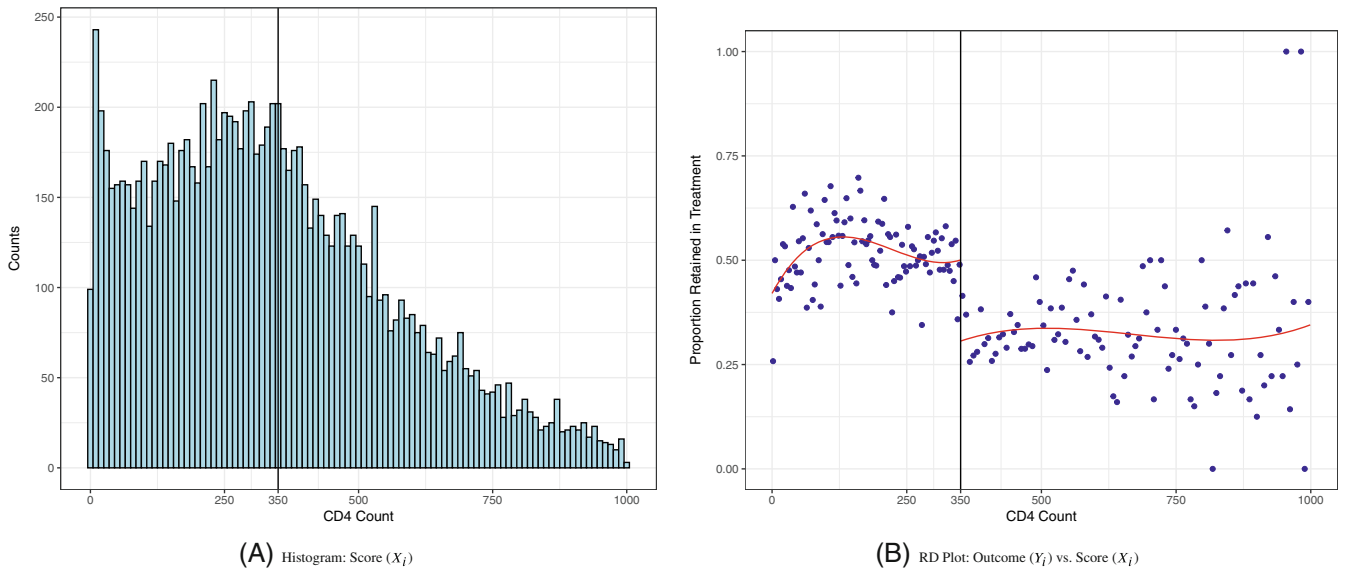


FIGURE 1 Basic plots—ART application. The score X_i is patient i's CD4 count, and all patients below the cutoff are assigned to receive ART. In panel (B), the outcome is Y_i, an indicator equal to 1 if patient i was retained in care (0 otherwise); dots are local means of Y_i calculated for patients in different nonoverlapping bins of X_i; and the solid line is a 4th-order polynomial of Y_i on X_i, fitted separately for patients above and below the cutoff.

Figure 1B presents a canonical RD plot of the observed outcome variable given the score.[20] Although we could construct a raw scatter plot of the outcome against the score, such plot is often be uninformative and hides many interesting features in the outcome-score relationship like discontinuities or nonlinearities. For this reason, it is customary to "smooth" the data before plotting, which is done by binning the support of the score into disjoint (ie, nonoverlapping) intervals, and then reporting the average outcome for units with score within each bin, an approach conceptually analogous to the histogram in Figure 1B. These binned means can be interpreted as a nonsmooth local approximation to the unknown regression functions of Y_i given X_i. The standard RD plot consists of these binned means with the addition of two global polynomials fits, one above and one below the cutoff, based on regressing the outcome Y_i on a polynomial of X_i using the raw (ie, not binned) data. The global polynomial fits can be interpreted as a smooth global approximation of the unknown regression functions, in contrast to the nonsmooth approximation provided by the local means. Choosing the appropriate global polynomial order is important: when the order of the polynomial is "too" high, the global polynomial regression will over-fit the data. This over-fitting is usually referred to as Runge's phenomenon, and is known to be particularly detrimental at boundary points, which is the area of interest in RD designs. See Cattaneo et al[15(section 2)] for more details, and Cattaneo et al[23] for related visualization methods in other empirical contexts.

The RD plot in Figure 1B gives a first glance at the RD design for the ART application. It shows that patients assigned to treatment (CD4 count strictly below 350) had an average retention higher than those assigned to control. In RD designs, however, identification of treatment effect occurs at or near the cutoff, where treatment assignment changes discontinuously but all other confounders are assumed to change smoothly or not at all. To formalize this intuition, we need to introduce key assumptions underlying RD designs and also define treatment effects (or parameters) of interest by "localizing" near the cutoff. Following the taxonomy introduced by Cattaneo et al,[24] we consider the continuity framework and the local randomization framework for the analysis and interpretation of RD designs in both sharp and fuzzy RD designs.

2.3 | Sharp RD designs

We first discuss settings with perfect treatment compliance. This is not the case for the ART application, or many other biomedical applications, but this simpler setup helps us put forth key concepts without added complications. The next section generalizes the setup to allow for imperfect compliance (ie, fuzzy RD designs). As discussed there, if

researchers are interested on intention-to-treat effects, then the sharp RD design is indeed the appropriate setup to consider even in the presence of noncompliance. As a result, the discussion in this section is a key building block for Fuzzy RD analysis.

We adopt the standard potential outcomes framework and assume that each unit has one outcome corresponding to each possible value of the treatment assignment: $Y_i(0)$ under control assignment, and $Y_i(1)$ under treatment assignment. The observed outcome is determined by the potential outcome corresponding to the treatment assigned to each unit: $Y_i = (1 - T_i) \cdot Y_i(0) + T_i \cdot Y_i(1)$. The observed data is $(Y_1, X_1), \ldots, (Y_n, X_n)$. In most of our discussion, we assume that the observations are a random sample with random potential outcomes. We deviate from this setup only when discussing analysis of experiments approaches based on Fisherian inference or Neyman methods, which assume that the potential outcomes are nonstochastic and hence that the observed outcomes are random only because of the randomness induced by the treatment assignment mechanism, that is, the probability distribution determining $(T_1, T_2, \ldots, T_n)$.

2.3.1 | Continuity-based framework

In the sharp RD design with continuous score, the leading conceptual approach is the continuity-based framework,[25] where the causal treatment effect is the average treatment effect at the cutoff:

$$\tau_{\text{SRD}} \equiv \mathbb{E}[Y_i(1) - Y_i(0) | X_i = c]. \tag{1}$$

In this framework, potential outcomes are always assumed to be random, so the conditional expectations are interpreted and computed in the usual way. The sharp RD treatment effect τ_{SRD} is the average effect of treatment for units *local* to the cutoff—that is, for units with score values $X_i = c$.

The identification of τ_{SRD} is based on the idea that units with similar values of the score but on opposite sides of the cutoff should be "comparable" in all predetermined characteristics except for the fact that units whose scores are above the cutoff are assigned to treatment while units whose scores are below the cutoff are not. Predetermined characteristics, also known as pretreatment or predetermined covariates, are all features of the units whose values are determined before the treatment is assigned. For example, in the ART example, the age and sex of patients are predetermined covariates.

We can formalize the logic of comparability at the cutoff using continuity, which relies on mild extrapolation for units with score near the cutoff. First, we define the average potential outcomes given the score: $\mathbb{E}[Y_i(1)|X_i = x]$ and $\mathbb{E}[Y_i(0)|X_i = x]$. These conditional expectation functions are usually called *regression functions*, and are unknown; the two solid lines in Figure 1B depict global polynomial approximations to these functions in the ART application. If the regression functions $\mathbb{E}[Y_i(1)|X_i = x]$ and $\mathbb{E}[Y_i(0)|X_i = x]$, seen as functions of x, are continuous at $x = c$, then the units will be comparable "just" above and below the cutoff. That is, under the assumption of continuity, we can use the regression functions to link observed data to counterfactual quantities in the following way:

$$\tau_{\text{SRD}} = \lim_{x \downarrow c} \mathbb{E}[Y_i | X_i = x] - \lim_{x \uparrow c} \mathbb{E}[Y_i | X_i = x]. \tag{2}$$

In Equation (2), continuity implies that as the score value gets closer to the cutoff c, the average potential outcome function $\mathbb{E}[Y_i(0)|X_i = x]$ gets closer to its value at the cutoff, $\mathbb{E}[Y_i(0)|X_i = c]$, and analogously for $\mathbb{E}[Y_i(1)|X_i = x]$. Thus, continuity gives a formal justification for estimating the sharp RD effect by focusing on observations in a small neighborhood above and below the cutoff to estimate, respectively and separately, $\mathbb{E}[Y_i(1)|X_i = c]$ and $\mathbb{E}[Y_i(0)|X_i = c]$. The observations in this neighborhood, by construction, will have similar score values; and by virtue of continuity, their average potential outcomes will also be similar. As mentioned above, employing global polynomial approximations should be avoided due to poor boundary behavior of such estimates; instead, the approximation should be local.

The logic of the continuity-based framework is graphically illustrated in Figure 2A. Continuity of the two conditional expectations ensures that the vertical distance between the two curves at c represents the RD estimand. We cannot directly estimate this quantity since we never observe the two curves at c: units with scores exactly at or just above c are treated, but units with scores just below c are control. Nevertheless, if the average potential outcomes at c are not abruptly different from the average potential outcomes at values of the score just below c, then units just above and below the cutoff should be comparable, and we can approximately identify the vertical distance at c using the local observed data, relying on minimal extrapolation in finite samples.

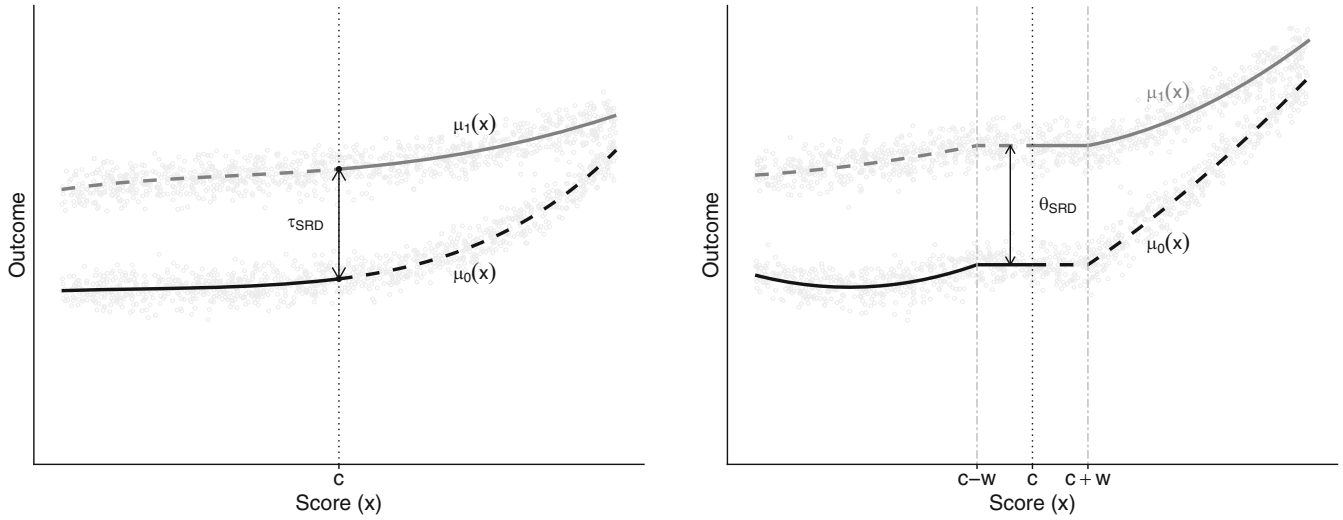


FIGURE 2 Graphical illustration of sharp RD design frameworks. $\mu_1(x)$ and $\mu_0(x)$ are the conditional expectation functions of the potential outcome under treatment and control, respectively, given the score—that is, $\mu_1(x) = \mathbb{E}[Y_i(1)|X_i = x]$ and $\mu_0(x) = \mathbb{E}[Y_i(0)|X_i = x]$. Dashed lines represent unobserved functions, solid lines represent observed functions.

The RD treatment effect τ_{SRD} differs from the two most common estimands often targeted in observational studies: the average treatment effect (ATE) and the average treatment effect on the treated (ATT). The ATE measures the average difference in outcomes when all individuals in the study population are assigned to treatment vs when all individuals are assigned to control. On the other hand, the ATT measures the average difference in outcomes among those individuals in the population that were actually exposed to the treatment. The RD estimand, however, is far more local than both of these estimands as it only applies to units close to the cutoff. Ideally, we would like to study more general treatment effects, such as the ATE, in order to learn about the average difference in outcomes that would occur if all units in the study were switched from treated to untreated. Unfortunately, this kind of treatment effects is not generally available in RD designs because the nonexperimental treatment assignment only justifies studying effects for units whose scores are near the cutoff—see Cattaneo et al[26] for further discussion and related references.

2.3.2 | Local randomization framework

The second framework for the analysis of RD designs is based on the idea of local randomization,[24,27] where potential outcomes could be viewed as random variables or as fixed quantities, exactly as in the analysis of experiments literature.[3-5] To formalize this framework, we introduce notation for the local randomization neighborhood or window, $\mathcal{W} = [c - w, c + w]$, where $w > 0$ is its half length and $\mathcal{W}$ is assumed symmetric around the cutoff only for simplicity. In this setting, we call $\mathcal{W}$ a *window* to distinguish it from the local neighborhood or bandwidth used in the context of continuity-based methods.

While in the continuity-based framework the key assumption is continuity of conditional expectations to enable extrapolation to the cutoff, in the local randomization framework the idea is to impose conditions to induce an experimental setting near the cutoff. Thus, the key two assumptions are: (i) known treatment assignment mechanism for all units with score in $\mathcal{W}$; and (ii) lack of relationship between score and outcomes for all units with score in $\mathcal{W}$. The second assumption is important. In the continuity-based RD design, the fact that the score is related to the potential outcomes does not present challenges because the parameter of interest is defined at the (single) cutoff point. In contrast, in the local randomization framework, the potential outcomes can be related to the score far from the cutoff, but this relationship must vanish in the window $\mathcal{W}$. As such, we must assume that the value of the score within this interval is unrelated to the potential outcomes—a condition that is not guaranteed by the random assignment of the score X_i, nor by the random assignment of the treatment T_i. Such an assumption is plausible for small neighborhoods around the cutoff, that is, for

those units that have scores closest to the cutoff. See Cattaneo et al[16(section 2)] and references therein for more discussion and extensions.

In the local randomization framework, we can define treatment effects that are analogous to those discussed in the continuity-based framework. The main difference is that the continuity-based estimands are defined at the cutoff, and the analogous local randomization estimands are defined in the window $\mathcal{W}$ around the cutoff. The local randomization sharp RD parameter is the average treatment effect inside the window $\mathcal{W}$, analogous to τ_{SRD}, defined as

$$\theta_{\text{SRD}} \equiv \mathbb{E}_{\mathcal{W}}[Y_i(1) - Y_i(0)] = \frac{1}{N_{\mathcal{W}}} \sum_{i:X_i \in \mathcal{W}} \mathbb{E}_{\mathcal{W}}\left[\frac{T_i Y_i}{\mathbb{P}_{\mathcal{W}}[T_i = 1]}\right] - \frac{1}{N_{\mathcal{W}}} \sum_{i:X_i \in \mathcal{W}} \mathbb{E}_{\mathcal{W}}\left[\frac{(1-T_i) Y_i}{1 - \mathbb{P}_{\mathcal{W}}[T_i = 1]}\right], \tag{3}$$

where the potential outcomes $Y_i(0)$ and $Y_i(1)$ can be taken as random or fixed depending on the approach taken, $\mathbb{P}_{\mathcal{W}}[\cdot]$ and $\mathbb{E}_{\mathcal{W}}[\cdot]$ denote the probability and expectation taken conditionally for those units with $X_i \in \mathcal{W}$, and $N_{\mathcal{W}}$ is the number of units with $X_i \in \mathcal{W}$. The last expression after the equality sign indicates that θ_{SRD} can be estimated from the data just like in the standard analysis of experiments, but using only units with score within the local randomization window $\mathcal{W}$.

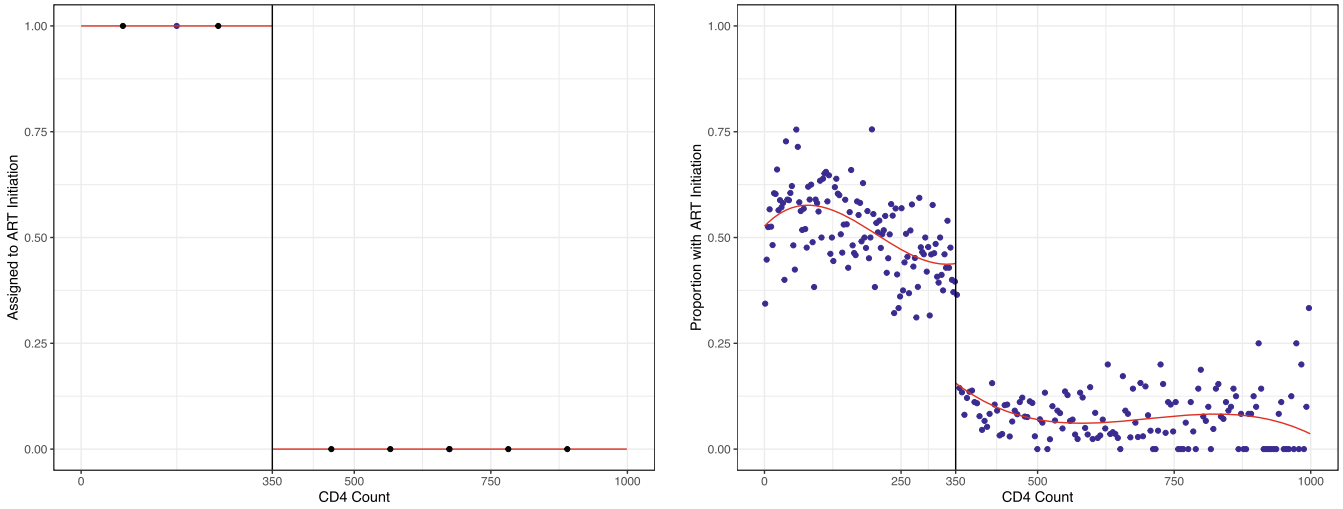
Figure 2B showcases the local randomization framework, showing a local neighborhood around c defined by $\mathcal{W}$. The key idea is that there exists a neighborhood or window around the cutoff where the treatment assignment resembles what it would have been in a randomized experiment. Given this fact, we can simply estimate the treatment effect as if this was an experiment for the units that fall within the local neighborhood around c. As we discuss below, the analogy between RD local randomization and a true experiment is not perfect, and the local randomization RD framework requires stronger assumptions than the continuity-based framework. However, local randomization methods are valid when the score is discrete, while continuity-based methods may be invalid if the score is too coarse (Section 4).

2.4 | Fuzzy RD designs

An important feature of the ART application, which is common in many RD designs in biomedical research, is that being *assigned* to the treatment condition is not the same as actually *receiving* the treatment. We use the binary variable D_i to denote whether the treatment is actually received by unit i ($D_i = 1$) or not ($D_i = 0$), while we continue to use the binary variable T_i to record whether the treatment is offered ($T_i = 1$) or not ($T_i = 0$). In the sharp RD design, we always have $T_i = D_i$ because compliance with treatment assignment is perfect, while the defining feature of a fuzzy RD design is that there are some units for which $T_i \neq D_i$. For example, in the ART application, there are patients with CD4 counts of less than 350 ($T_i = \mathbb{1}(X_i < 350) = 1$) who never initiate ART ($D_i = 0$). This is depicted visually in Figure 3 using RD plots. Figure 3A shows that all patients with $T_i = \mathbb{1}(X_i < 350) = 1$ where assigned to treatment with probability one, while all patients with $T_i = 0$ where assigned to control with probability one. However, treatment assignment was not always followed, as shown in Figure 3B, which plots the proportion of patients actually receiving ART against the score.

We employ potential outcomes to formalize fuzzy RD designs. Every unit has two *potential treatments*: $D_i(1)$ is the treatment that unit i receives when this unit is assigned to the treatment condition (ie, when $T_i = 1$), while $D_i(0)$ is the treatment that unit i receives when this unit is assigned to the control condition (ie, when $T_i = 0$). Both $D_i(1)$ and $D_i(0)$ can be one or zero, depending on unit i's compliance decisions. For example, if unit i is assigned to the treatment condition but refuses to receive the treatment, $D_i(1) = 0$; and a unit that complies perfectly with their assignment has $D_i(1) = 1$ and $D_i(0) = 0$. Thus, the observed treatment received is $D_i = (1 - T_i) \cdot D_i(0) + T_i \cdot D_i(1)$.

For the outcome of interest, this framework implies that every unit has four different potential outcomes depending on the combination of treatment assignment and compliance decisions: $Y_i(1,0)$, $Y_i(1,1)$, $Y_i(0,0)$, and $Y_i(0,1)$. We denote them generally as $Y_i(T_i, D_i(T_i))$, a function of both the treatment assigned and the treatment received. For example, $Y_i(0,1)$ corresponds to the potential outcome that would occur if unit i were assigned to the control condition ($T_i = 0$) but received the treatment anyway ($D_i(0) = 1$). However, we only observe the potential outcome and the potential treatment corresponding to the values of T_i and D_i that are realized for unit i. Formally, the observed outcome is now $Y_i = (1 - T_i) \cdot Y_i(0, D_i(0)) + T_i \cdot Y_i(1, D_i(1))$, and the observed data is $(Y_1, D_1, X_1), \ldots, (Y_n, D_n, X_n)$.



(A) Treatment Assignment ($\mathbb{P}[T_i = 1|X_i]$) (B) Treatment Take-up ($\mathbb{P}[D_i = 1|X_i]$)

FIGURE 3 Treatment assignment vs treatment take-up—ART application. The score X_i is patient i's CD4 count, and all patients below the cutoff are assigned to receive ART. In panel (A) the y-axis is D_i, an indicator equal to 1 if patient i received ART (0 otherwise); the dots are local means of D_i calculated for patients in different nonoverlapping bins of X_i, separately above and below the cutoff. In panel (B), the y-axis is Y_i, an indicator equal to one if patient i was retained in case (0 otherwise); the dots are local means of Y_i calculated for patients in different nonoverlapping bins of X_i; and the solid line is a 4th-order polynomial of Y_i on X_i, fitted separately for patients above and below the cutoff.

2.4.1 | Continuity-based framework

In the fuzzy RD design, the standard RD estimand, τ_{SRD}, is unavailable except under strong assumptions that will be implausible in many applications (eg, constant treatment effects as a function of the score). Instead, when there is noncompliance, researchers typically focus on two types of treatment effects: the effects of *assigning* the treatment for all units, and the effect of *receiving* the treatment for a subpopulation of units. Each type of effect requires different assumptions, and which one is of interest depends on the particular application.

The effect of the treatment received is of obvious importance. For example, in the ART application we are interested in the effect of initiating ART on patient retention. However, in some cases, researchers are also interested in the effect of assigning the treatment on the outcome, which is commonly known as the *intention-to-treat* (ITT) effect. This effect includes not only the effect that the treatment received may directly have on the outcome, but also the effect caused by strategic compliance decisions that individuals make in response to knowledge about their assignment. Policy-makers interested in anticipating the overall effects of establishing a new program are often interested in ITT effects.

Within the continuity-based framework, we start by considering the effect of treatment assignment on the outcome (Y_i) and on the treatment received (D_i), both of which can be seen as sharp RD effects of the treatment assignment. The RD effect of the treatment assignment on the observed outcome is

$$\tau_{\text{Y}} \equiv \lim_{x \downarrow c} \mathbb{E}[Y_i|X_i = x] - \lim_{x \uparrow c} \mathbb{E}[Y_i|X_i = x]. \tag{4}$$

Under continuity assumptions analogous to those in the canonical sharp RD case, τ_{Y} captures the ITT effect of the treatment assignment on the outcome at the cutoff, which we can write as $\tau_{\text{Y}} = \mathbb{E}[Y_i(1, D_i(1)) - Y_i(0, D_i(0))|X_i = c]$, the average change in potential outcomes at the cutoff from switching the assignment from control to treated. In the ART application, this effect is plotted in Figure 1B using global polynomial approximations. The ITT effect of the treatment assignment on the outcome follows a sharp RD design where the T_i is seen as the treatment of interest. Thus, we estimate the same difference in limits $\lim_{x \downarrow c} \mathbb{E}[Y_i|X_i = x] - \lim_{x \uparrow c} \mathbb{E}[Y_i|X_i = x]$ that we estimate in a sharp RD setting, but we modify the assumptions and interpretation to accommodate imperfect compliance. Because some units fail to comply with their assignment, the sharp RD treatment effect of T_i on Y_i is no longer the effect of the treatment itself, but rather the effect of *assigning* the treatment. For example, in the ART application, τ_{Y} captures the average effect of offering ART to patients

whose CD4 is 350 who may or may not accept the offer, while the parameter τ_{SRD} would capture the effect of actually starting ART for those patients.

The RD effect of the treatment assignment on the treatment received at the cutoff is

$$\tau_{\text{D}} \equiv \lim_{x \downarrow c} \mathbb{E}[D_i | X_i = x] - \lim_{x \uparrow c} \mathbb{E}[D_i | X_i = x]. \tag{5}$$

Since D_i is binary, τ_{D} captures the difference in the probability of receiving the treatment at the cutoff between units just assigned to the treatment vs assigned to the control condition. In the ART application, this is the difference between the proportion of patients with CD4 counts just below 350 who initiate ART and the proportion of patients with CD4 counts just above 350 who initiate ART. This treatment effect is illustrated in Figure 3B. Under continuity conditions, the difference in treatment probabilities captured by τ_{D} can be attributed to the RD assignment rule; in this case, τ_{D} represents the average effect of assigning the treatment on receiving the treatment at the cutoff, that is, $\tau_{\text{D}} = \mathbb{E}[D_i(1) - D_i(0) | X_i = c]$. This effect is usually called the *first-stage* or *take-up* effect. Both τ_{Y} and τ_{D} are sharp RD parameters.

Investigators are often also interested in the effect of receiving the treatment, not merely of assigning it. While τ_{SRD} is infeasible in the fuzzy RD design due to noncompliance, under additional assumptions, it is possible to estimate a related parameter that captures the average effect of the treatment at the cutoff for a particular subpopulation of units. We define the fuzzy RD treatment effect as

$$\tau_{\text{FRD}} \equiv \frac{\tau_{\text{Y}}}{\tau_{\text{D}}}, \tag{6}$$

which is the ratio of the sharp RD effect of T_i on Y_i and the sharp RD effect of T_i on D_i.

The parameter τ_{FRD} can be interpreted as the average effect of the treatment received at the cutoff for the subpopulation of units who are compliers—informally defined as units who receive the treatment when their score is above the cutoff and refuse the treatment when their score is below the cutoff. Different authors have formalized the definition of compliers differently in RD settings; a thorough discussion is beyond the scope of our discussion, but we refer the reader to References 28-30 for examples, and to References 16 (section 3) for a practical discussion. See also Baiocchi et al[31] for a review of IV methods for causal inference.

Regardless of the technical details, interpreting the fuzzy RD parameter as the effect of D_i on Y_i for the compliers requires three assumptions. The formalization of these assumptions varies depending on the particular definitions adopted, but the conceptual ideas are similar in all cases. First, the parameter τ_{D} must be nonzero, and well-separated from zero for estimation and inference to be meaningful. In other words, being above vs below the cutoff must induce some units to actually take the treatment. This rules out, for example, a situation where having a CD4 count below 350 induces no patients to start ART. This is usually referred to as the *relevance* assumption or the *first-stage* in IV settings, and is testable. We will showcase this point in Section 3.

Second, we need continuity conditions similar to those invoked in the sharp RD case, but generalized for the more complex setting of noncompliance. These continuity conditions will implicitly require, among other things, that the treatment assignment only affect the average outcomes via its effect on the treatment received, but not directly, analogous to the *exclusion* restriction in IV settings. In other words, T_i should only have an affect on Y_i through D_i: crossing the cutoff should only affect the outcome if it has an effect of changing the actual treatment received, but not otherwise. This key assumption is untestable and requires careful qualitative reasoning for justification, particularly in medical settings where placebo effects are common.[32]

In the ART application, the exclusion restriction requires that having a CD4 count below 350 have no effect on retention in care except by inducing people to initiate ART. This assumption might be implausible if seeing a CD4 count below 350 leads physicians to order additional tests or to communicate with patients differently, which can in turn lead to discovery of other health issues and return for care of a different condition. The exclusion restriction would be more plausible if the outcome were a biological manifestation of HIV rather than retention in care, as it is more plausible that the only way future HIV symptoms would be reduced is through exposure to ART.

Finally, it is also common to assume *monotonicity* or a similar condition for the interpretation of the fuzzy RD treatment effect τ_{FRD}. Informally, monotonicity requires that a patient who decides to receive the treatment when they are not eligible for it, continues to take the treatment when they are eligible. One way to interpret this in the RD setting where $T_i = \mathbb{1}(X_i \geq c)$, is to require that a unit with score X_i who refuses the treatment when the cutoff is c must also refuse the treatment for any cutoff $c' > c$, and a unit who takes the treatment when the cutoff is c must also take the treatment for

any cutoff $c' < c$. In our example, this implies that a patient who, say, has a CD4 count of 340 and refuses ART when the cutoff is 350, he or she must also refuse ART when the cutoff is 330.

2.4.2 | Local randomization framework

In the local randomization framework for fuzzy RD designs, we can consider parameters of interest that are analogous to those discussed in the continuity-based framework. The sharp RD estimator of the effect of T_i on Y_i and the effect of T_i on D_i are defined, respectively, as

$$\theta_{\mathrm{Y}} \equiv \frac{1}{N_{\mathcal{W}}} \sum_{i:X_i \in \mathcal{W}} \mathbb{E}_{\mathcal{W}}\left[\frac{T_i Y_i}{\mathbb{P}_{\mathcal{W}}[T_i = 1]}\right] - \frac{1}{N_{\mathcal{W}}} \sum_{i:X_i \in \mathcal{W}} \mathbb{E}_{\mathcal{W}}\left[\frac{(1-T_i) Y_i}{1 - \mathbb{P}_{\mathcal{W}}[T_i = 1]}\right] \tag{7}$$

and

$$\theta_{\mathrm{D}} \equiv \frac{1}{N_{\mathcal{W}}} \sum_{i:X_i \in \mathcal{W}} \mathbb{E}_{\mathcal{W}}\left[\frac{T_i D_i}{\mathbb{P}_{\mathcal{W}}[T_i = 1]}\right] - \frac{1}{N_{\mathcal{W}}} \sum_{i:X_i \in \mathcal{W}} \mathbb{E}_{\mathcal{W}}\left[\frac{(1-T_i) D_i}{1 - \mathbb{P}_{\mathcal{W}}[T_i = 1]}\right], \tag{8}$$

which parallel the continuity-based parameters τ_{Y} and τ_{D}. Under the local randomization assumptions, the parameters θ_{Y} and θ_{D} capture the average effect of assigning the treatment for observations with scores in the window. Finally, we can also define the local-randomization fuzzy RD parameter as the ratio: $\theta_{\mathrm{FRD}} \equiv \theta_{\mathrm{Y}}/\theta_{\mathrm{D}}$.

As in the continuity-based framework, under appropriate assumptions, θ_{FRD} can be interpreted as the average treatment effect in the window for compliers. The assumptions typically used are similar to those required in IV settings, now applied to observations with scores in the window $\mathcal{W}$, and hence similar to those discussed for the continuity-based framework. Once again, the effect of the treatment assignment on the treatment received, θ_{D}, must be well separated from zero. The exclusion restriction that the treatment assignment have no direct effect on the outcomes must also hold for all units with scores within the window; this restriction is implied by the local randomization condition that the (distribution of the) potential outcomes and potential treatments is not a function of the score inside $\mathcal{W}$. Finally, the assumption of monotonicity requires that there be no units with scores in $\mathcal{W}$ who receive a treatment condition that is always opposite to their assignment.

Finally, it is important to understand how to interpret the fuzzy RD estimands τ_{FRD} and θ_{FRD}. These estimands differ from the (local) average treatment effects that are commonly used in the IV literature:[31] the fuzzy RD estimands capture the average treatment effect for a subpopulation (eg, compliers) with a score value at or near the cutoff, and by implication often have lower external validity than the standard IV estimand. In the ART application, the fuzzy RD treatment effects only apply to the set of compliers with scores near 350. The fuzzy RD treatment effect may differ compared to those patients with much higher or lower CD4 counts. For more discussion on extrapolation of RD treatment effects away from the cutoff, see Reference 26 and references therein.

3 | ANALYSIS WITH CONTINUOUS SCORE

We now discuss estimation, inference, and validation methods within the continuity-based and the local randomization RD frameworks with a continuously distributed score, using again the ART application as the running empirical example. All the results in this section can be reproduced using the replication materials.

3.1 | Continuity-based methods

A common problem in RD settings is that there are often few observations with score values very close to the cutoff, which means that estimating the effect at $X_i = c$ requires using observations whose values of X_i are relatively far from c. Because a sufficiently smooth function can be well approximated by a polynomial function, up to misspecification error, standard continuity-based RD estimation methods approximate the regression functions, $\mathbb{E}[Y_i(0)|X_i = x]$ and $\mathbb{E}[Y_i(1)|X_i = x]$ using a polynomial function of the score.

3.1.1 | Point estimation

Modern RD estimation is based on local polynomial approximations that discard observations sufficiently far away from the cutoff and then employ a low-order polynomial approximation (usually linear or quadratic) for estimation. This approach is known as local polynomial regression in the statistical literature.[33] State-of-the-art RD methods use two separate linear polynomial fits for treated and control units using only observations near the cutoff as determined by the choice of a bandwidth parameter. This local approach is more robust and less sensitive to boundary and over-fitting problems.

Local polynomial methods require the user to make three choices: the bandwidth, the kernel function, and the polynomial order. The bandwidth controls the width of the neighborhood around the cutoff that is used to fit the local polynomial models, and hence determines the number of observations above and below the cutoff that are used for estimation. Within the neighborhood determined by the bandwidth, it is common to adopt a weighting scheme to ensure that the observations closer to c receive more weight than those further away. The weighting scheme is referred to as a kernel function, $K(\cdot)$, and two common options are the triangular kernel, $K(x) = (1 - |x|)\mathbb{1}(|x| \leq 1)$, which linearly down-weights observations within the bandwidth, and the uniform kernel, $K(x) = \mathbb{1}(|x| \leq 1)$, which gives equal weight to all observations within the bandwidth. The polynomial order p determines the order of the polynomial approximation near the cutoff. In the software resources used in this tutorial, the defaults are linear fit ($p = 1$) and triangular kernel. These choices have objective theoretical advantages in the nonparametrics literature,[33] but the researcher can also investigate the robustness of the empirical results by choosing $p = 2$ or a uniform kernel.

The RD estimate is thus constructed as follows. For observations above the cutoff (ie, observations with $X_i \geq c$), fit a weighted least squares regression of the outcome Y_i on a constant and $(X_i - c), (X_i - c)^2, \ldots, (X_i - c)^p$ with weight $K\left(\frac{X_i - c}{h}\right)$ for each observation, leading to the estimated equation $\widehat{Y}_i = \widehat{\mu}_+ + \widehat{\mu}_{+,1}(X_i - c) + \widehat{\mu}_{+,2}(X_i - c)^2 + \cdots + \widehat{\mu}_{+,p}(X_i - c)^p$, where the estimated intercept, $\widehat{\mu}_+$, is a point estimate of $\mu_+ = \mathbb{E}[Y_i(1)|X_i = c]$. Similarly, for observations below the cutoff, fit a weighted least squares regression of the outcome Y_i on a constant and $(X_i - c), (X_i - c)^2, \ldots, (X_i - c)^p$ with weight $K\left(\frac{X_i - c}{h}\right)$ for each observation, leading to $\widehat{Y}_i = \widehat{\mu}_- + \widehat{\mu}_{-,1}(X_i - c) + \widehat{\mu}_{-,2}(X_i - c)^2 + \cdots + \widehat{\mu}_{-,p}(X_i - c)^p$, where the estimated intercept, $\widehat{\mu}_-$, is a point estimate of $\mu_- = \mathbb{E}[Y_i(0)|X_i = c]$. Therefore, the sharp RD point estimate is

$$\widehat{\tau}_{\mathtt{SRD}} = \widehat{\mu}_+ - \widehat{\mu}_-. \tag{9}$$

The choice of bandwidth h, which determines which observations near the cutoff are used, is the most critical when implementing local polynomial RD methods. A small bandwidth will reduce the approximation error of the local polynomial approximation because it only uses observations very close to the cutoff. However, a small bandwidth will also increase the variance of the estimates because only a few observations are used in the local fit. Analogously, a large bandwidth may increase the approximation error if the underlying regression function differs considerably from the polynomial approximation used, but will result in lower variance due to the relatively larger number of observations included. Thus, bandwidth selection embodies a bias-variance trade-off: smaller bandwidths will tend to have less bias but higher variance, and viceversa. The mean squared error (MSE) of any estimator is the sum of its bias squared plus its variance; given the bias-variance tradeoff, bandwidth selection can be automated in a principled, data-driven way by first deriving an approximation to the MSE of the RD point estimator, and then choosing the value of h that minimizes it. This so-called MSE-optimal bandwidth selection approach has become the standard for RD estimates. See Calonico et al[34,35] for the most recent methodological developments, and Cattaneo et al[36] and Calonico et al[37] for an overview on neighborhood selection methods in RD designs more generally.

3.1.2 | Confidence intervals

The MSE-optimal bandwidth is used to construct an MSE-optimal point estimator, $\widehat{\tau}_{\mathtt{SRD}}$, but using that bandwidth to conduct standard least squares inference is in general *invalid*.[34,35,37] To be more precise, the MSE-optimal bandwidth balances bias and variance in such a way that the point RD estimator exhibits a misspecification bias in its distribution, which leads to confidence intervals and hypothesis tests that are invalid in general, even in large samples. This implies that the usual asymptotic 95-percent confidence interval for $\tau_{\mathtt{SRD}}$ given by $\mathtt{CI} = \left[\widehat{\tau}_{\mathtt{SRD}} \pm 1.96 \cdot \sqrt{\widehat{\mathbb{V}}}\right]$, where $\widehat{\mathbb{V}}$ denotes a variance estimator, is invalid because the underlying Gaussian distribution of the RD point estimator has a

nonzero bias when the MSE-optimal bandwidth is used. It can be shown that CI will cover the population treatment effect τ_{SRD} roughly 80% of the time in repeated sampling, implying a false rejection rate of about 15-percentage points.

A principled alternative is to use the *robust bias corrected* confidence intervals proposed by Calonico et al,[34] and later extended to other settings.[35,38-41] Robust bias corrected confidence intervals modify the classical confidence intervals CI in two ways: (i) the point estimator $\hat{\tau}_{\text{SRD}}$ is debiased by including an estimate of the leading misspecification error (denoted by $\hat{\text{B}}$), and (ii) the variance estimator $\hat{\text{V}}$ is increased to incorporate the contribution of the bias correction step to the overall variability of the confidence interval (denoted by $\hat{\text{W}}$). Thus, the robust bias corrected confidence intervals take the form

$$\text{CI}_{\text{RBC}} = \left[\left(\hat{\tau}_{\text{SRD}} - \hat{\text{B}}\right) \pm 1.96 \cdot \sqrt{\hat{\text{V}} + \hat{\text{W}}} \right].$$

These confidence intervals are valid even when the MSE-optimal bandwidth is used, and have several demonstrable theoretical properties, including smaller coverage errors and less sensitivity to tuning parameter choices.[42-45] Furthermore, the improved finite sample performance of these intervals has been validated empirically.[46-48]

Our practical recommendation is therefore to (i) report the MSE-optimal RD point estimate $\hat{\tau}_{\text{SRD}}$, which is constructed using an MSE-optimal bandwidth choice, and (ii) report robust bias corrected confidence intervals, which employ the same MSE-optimal bandwidth choice. All these methods are readily available in Python, R, and Stata general-purpose software packages (https://rdpackages.github.io/). We use these methods for the analysis of our three empirical examples, as illustrated in the accompanying replication files.

The local polynomial methods for sharp RD continuity-based analysis can be extended to fuzzy RD designs to estimate τ_{Y}, τ_{D}, and τ_{FRD}. The first point estimator is exactly the same as described above, using local polynomials to estimate the relationship between Y_i and the score X_i—that is, $\hat{\tau}_{\text{Y}} = \hat{\tau}_{\text{SRD}}$. The estimator $\hat{\tau}_{\text{D}}$ of τ_{D} is constructed analogously, after replacing the observed outcome variable Y_i with the observed treatment status D_i. Once $\hat{\tau}_{\text{Y}}$ and $\hat{\tau}_{\text{D}}$ are available, the fuzzy RD estimand τ_{FRD} is estimated using $\hat{\tau}_{\text{FRD}} = \hat{\tau}_{\text{Y}}/\hat{\tau}_{\text{D}}$.

The estimator $\hat{\tau}_{\text{FRD}}$ is consistent for τ_{FRD} under standard regularity conditions, although it may exhibit more bias or other potential problems due to its intrinsic ratio structure. Heuristically, everything discussed in this section still applies to this estimator, but some more details are necessary. First, bandwidth selection can still proceed based on a MSE approximation, although now such approximation should also take into account the ratio structure of the estimator. Furthermore, more than one natural MSE-optimal bandwidth choice is available: it is possible to consider one single bandwidth for the ratio $\hat{\tau}_{\text{FRD}}$, or two distinct bandwidth choices, one each for the numerator and denominator. In practice, most researchers employ a single MSE-optimal choice for $\hat{\tau}_{\text{FRD}}$ or for $\hat{\tau}_{\text{Y}} = \hat{\tau}_{\text{SRD}}$, although some researchers prefer to choose two different bandwidths for $\hat{\tau}_{\text{Y}}$ and $\hat{\tau}_{\text{D}}$. As a general rule, it is usually recommended to use a single MSE-optimal bandwidth for the estimator of interest, in this case, $\hat{\tau}_{\text{FRD}}$. For inference, the same problems of misspecification biases arise in the fuzzy RD design, usually made more acute by the ratio structure of the point estimator. As a consequence, robust bias correction continues to be recommended whenever an MSE-optimal bandwidth choice is used for point estimation. Because these formulas are cumbersome we do not reproduce them here, but they can all be found in Calonico et al.[34,35,37]

3.1.3 | Continuity-based analysis of ART example

We now illustrate all the methods discussed so far using the ART application. The effects on the main outcome of interest are reported in Table 1. All the results in this table can be generated using rdrobust in any of the three software platforms (Python, R, Stata). First, we focus on the effect of being assigned to treatment (in this case, having a score below the cutoff). We find that having a CD4 count of 350 or greater reduces the likelihood of ART initiation by 21 percentage points ($\hat{\tau}_{\text{D}}$) and also reduces program retention by 14 percentage points ($\hat{\tau}_{\text{Y}}$). This means that being just below the 350 threshold increases likelihood of both ART and program retention. These are the effects of assignment to ART rather than of actual ART initiation, and as such do not fully capture the primary effect of interest—the effect of ART initiation on program retention. To explore the latter effect, we focus on the fuzzy RD estimate, $\hat{\tau}_{\text{FRD}}$, which is simply the ratio of the two ITT effects. We find that ART initiation increases program retention by more than 67 percentage points, and the confidence interval is bounded away from zero. Thus, we conclude that patients who initiated ART were much more likely to be retained in the treatment program. Under standard fuzzy RD assumptions, this is the effect on program

TABLE 1 RD estimation and Inference— ART application.

Continuity-based methods					
	RD Effect	**95% Robust CI**	**Bandwidth (h)**	N_h^-	N_h^+
ITT effect of ART assignment on ART initiation	−0.21	[−0.28, −0.12]	114.36	1494	1188
ITT effect of ART assignment on program retention	−0.14	[−0.22, −0.05]	114.36	1494	1188
Fuzzy effect of ART initiation on program retention	0.67	[0.34,1]	114.36	1494	1188
Local randomization methods					
	Risk Difference	**95% Confidence interval**	**Window ($\mathcal{W}$)**	$N_{\mathcal{W}}^-$	$N_{\mathcal{W}}^+$
ITT Effect of ART assignment on ART initiation	0.02	[−0.15, 0.19]	[346,354]	62	58
ITT Effect of ART assignment on program retention	−0.02	[−0.2, 0.16]	[346,354]	62	58
Fuzzy effect of ART initiation on program retention	−0.8	[−12.5, 10.91]	[346,354]	62	58

Note: The first three rows show, respectively, $\widehat{\tau}_{\text{D}}$, $\widehat{\tau}_{\text{Y}}$, and $\widehat{\tau}_{\text{FRD}}$, corresponding to the continuity-based estimates based on local linear estimation with MSE-optimal main bandwidth reported in third column. Column labeled "95% Robust CI" reports the robust 95% confidence intervals based on robust bias-corrected inference. Column N_h^- reports the number of observations with score in $[c-h, c)$ and column N_h^+ reports the number of observations with score in $[c, c+h]$. The last three show, respectively, $\widehat{\theta}_{\text{D}}$, $\widehat{\theta}_{\text{Y}}$, and $\widehat{\theta}_{\text{FRD}}$, corresponding to the local randomization estimates based on the data-driven chosen local randomization window reported in third column. Column $N_{\mathcal{W}}^-$ reports the number of observations with score in $\mathcal{W}$ and below the cutoff ($T_i = 0$) and column $N_{\mathcal{W}}^+$ reports the number of observations with score in $\mathcal{W}$ and above the cutoff ($T_i = 1$).

retention of initiating ART for patients with a CD4 count of 350 who are compliers. Recall that this interpretation requires, among other assumptions, that there is no effect of having a CD4 count below 350 on patient retention except via ART initiation.

We note that the analysis of RD designs can be enhanced by including predetermined covariates, which can be incorporated in a variety of ways. As in randomized experiments, a natural use of covariates is to improve the efficiency of the local polynomial RD estimator, as developed in Calonico et al.[35] However, predetermined covariates cannot be used to salvage an invalid RD design because incorporating covariates on those settings necessarily changes the RD parameters interest. See Cattaneo et al[49] for more discussion and references.

3.2 | Local randomization methods

The practical implementation of the local randomization framework requires two steps: (i) choosing the window $\mathcal{W}$ where the local randomization conditions are assumed to hold, and (ii) deploying methods from the analysis of experiments to perform estimation and inference for observations whose scores are inside the window.

Window selection is the most important step in the implementation of the local randomization approach for RD analysis. Although $\mathcal{W}$ could be selected in an ad-hoc fashion, a more principled approach is to select it using predetermined covariates, as proposed by Catteneo et al.[27] See also Catteneo et al[24] and Catteneo et al[16](section 2).

This data-driven window selection method requires that there be a set of predetermined covariates, Z, that are related to the score everywhere except inside $\mathcal{W}$. Once these predetermined covariates are chosen, the implementation of the window selection can be based on methods that assume random sampling (usually called super-population methods) or methods that condition on the units in the sample and assume that the only randomness comes from the treatment assignment mechanism (called Fisherian methods after statistician Ronald Fisher). For an in-depth review of super-population vs Fisherian methods in the causal inference framework, see Rosenbaum[4] and Imbens and Rubin.[5] Unlike super-population methods, Fisherian inference methods are exact in finite samples. Thus, in the context of RD window selection, it is often preferable to employ Fisherian methods because the windows considered typically have very few observations, which can invalidate the use of large-sample approximations.

For implementation, the researcher chooses a test statistic and performs a sequence of hypothesis tests that test the null hypothesis that the treatment has no effect on the covariates inside the window. The first test is conducted in the smallest window around the cutoff that has enough observations (typically a minimum of at least 10 observations on either side

is recommended); the sequence continues testing the null hypothesis of no treatment effect on Z in progressively larger windows until this hypothesis is rejected. While clearly this methodology relies on multiple hypothesis testing, there is no need to adjust the inferences because over-rejection of the null hypothesis leads to a more conservative window choice (ie, a smaller one). Consequently, a recommended rule is to reject all windows leading to p-values smaller than 0.15 or 0.10—these recommendations are based on power calculations under specific assumptions. (When Z includes multiple covariates, researchers can use the single p-value from an omnibus balance test, or the minimum p-value across individual balance tests.) The chosen $\mathcal{W}$ is the largest (symmetric) interval around the cutoff such that the predetermined covariates of the units inside the window are balanced between treated and control in that window, and in all smaller windows contained in it.

Once window selection is complete, analysis within the local randomization framework is straightforward. Under super-population methods, for example, a natural estimator for $\theta_{\texttt{SRD}}$ is the difference in means between the observed outcomes in the treated and control groups. When compliance is imperfect, we can estimate the sharp RD effects of T_i on Y_i and D_i, $\theta_{\texttt{Y}}$, and $\theta_{\texttt{D}}$, with the difference in the average observed outcomes between the treated and control groups inside the window, denoted by $\widehat{\theta}_{\texttt{Y}}$ and $\widehat{\theta}_{\texttt{D}}$, respectively. We can then estimate the local randomization fuzzy RD parameter, $\theta_{\texttt{FRD}}$, with $\widehat{\theta}_{\texttt{FRD}} = \widehat{\theta}_{\texttt{Y}}/\widehat{\theta}_{\texttt{D}}$. In the super-population framework, statistical inferences are based on standard large-sample approximations. In the specific context of RD, this means that the number of units within $\mathcal{W}$ is assumed to be large enough for distributional approximations to hold. This approach directly justifies the use of confidence intervals and p-values based on the large-sample properties of common test statistics such as standardized difference-in-means, least-squares and two-stage least-squares coefficients, and so forth, frequently used in the analysis of experiments.

When the number of observations in $\mathcal{W}$ is small, adopting a Fisherian approach is more appropriate. This approach takes potential outcomes as nonrandom and assumes that the randomization mechanism that assigned units to treated and control is either known or can be approximated. The fixed-margins assignment assumption is a natural choice. Fisherian randomization inference employs the sharp null hypothesis of no treatment effect for any unit within $\mathcal{W}$, controlling Type I error for any sample size. Most applications employ the difference-in-means between control and treatment units as the test statistics, but other choices are possible. Under additional assumptions on the treatment effect structure, point estimators and confidence intervals can also be constructed. For example, if we assume $Y_i(1) = Y_i(0) + \tau$, called a constant treatment effect model, we can form a point estimator and confidence intervals for τ based on standard Fisherian methods. Fisherian methods are also available for fuzzy RD designs where compliance is imperfect. See Ernst[50] for a review on permutation-based methods, Catteno et al[16,24,27] for more details on local randomization RD analysis, and Keele et al[51] and Kang et al[52] for related methodological developments for IV designs, which could be developed in the context of RD designs.

We illustrate the local randomization RD approach with the ART application. The analysis begins with window selection. The `rdlocrand` package contains tailored functions for the analysis of RD designs under local randomization, including a function for window selection. Using the data-driven methods for window selection outlined above, the window selected is [346,354]—that is, we find that predetermined covariates in the data are balanced for patients with CD4 counts between 346 and 354. We omit the balance test results for space considerations, but the window selection was based on the same covariates shown in Table 2. The resulting local randomization window is much narrower than the neighborhood implied by the bandwidth estimated using continuity-based methods, which is a common phenomenon in practice. The local randomization approach leads to a local neighborhood $\mathcal{W} =$ [346,354] with 121 patients, while the estimated bandwidth from the continuity-based analysis in Table 1 leads to the local region [239,461] with 2,593 patients. By their very nature, local randomization methods focus on much smaller neighborhoods around the cutoff, and thus use substantially fewer observations, which implies that they generally have less statistical precision than continuity-based methods. Nevertheless, local randomization methods can offer a useful complement and a robustness check for continuity-based methods when both frameworks are applicable.

Table 1 presents the main empirical results for the ART application using local randomization methods. Given the small sample sizes, all estimated effects are statistically indistinguishable from zero at conventional levels. But the confidence intervals do cover the point estimates reported with the continuity-based methods in Table 1. Thus, the results are statistically consistent with each other, albeit the local randomization methods are less informative than the continuity-based methods. One way to further investigate the role of lack of statistical precision is to increase the local randomization neighborhood. We regard this approach as a sensitivity test for the RD design, so we discuss it further below along with other falsification methods. Table 4 reports results for three wider windows—[340,360], [335,365], and

TABLE 2 Continuity-Based ITT RD estimates for predetermined covariates with robust bias corrected inference—ART application.

	Mean below	Mean above	$\widehat{\tau}_{\text{Y}}$	Robust *p*-value	MSE-optimal bandwidth	N_h^-	N_h^+
Age 0-18	0.07	0.08	0.01	0.44	126.62	2389	1893
Age 18-25	0.27	0.30	0.03	0.53	116.23	2178	1759
Age 25-30	0.24	0.19	−0.04	0.16	153.19	2860	2223
Age 30-35	0.14	0.14	−0.01	0.84	109.77	2056	1653
Age 35-40	0.09	0.10	0.01	0.69	143.56	2689	2102
Age 40-45	0.07	0.07	−0.00	0.93	106.65	1992	1617
Age 45-55	0.10	0.09	−0.00	0.81	131.88	2463	1953
Age 55+	0.04	0.03	−0.01	0.50	92.74	1733	1457
2011 Qtr3	0.13	0.11	−0.03	0.23	139.49	2666	2092
2011 Qtr4	0.18	0.17	−0.01	0.68	108.15	2085	1668
2012 Qtr1	0.19	0.21	0.02	0.49	158.60	2982	2330
2012 Qtr2	0.18	0.18	0.00	0.93	108.14	2085	1668
2012 Qtr3	0.18	0.19	0.01	0.58	130.95	2494	1974
2012 Qtr4	0.14	0.14	0.00	0.93	137.76	2632	2067
Female	0.68	0.73	0.05	0.18	89.75	1698	1422
Clinic A	0.17	0.13	−0.04	0.07	100.46	1916	1579
Clinic B	0.12	0.15	0.03	0.21	108.18	2085	1668
Clinic C	0.14	0.15	0.01	0.69	114.85	2182	1755

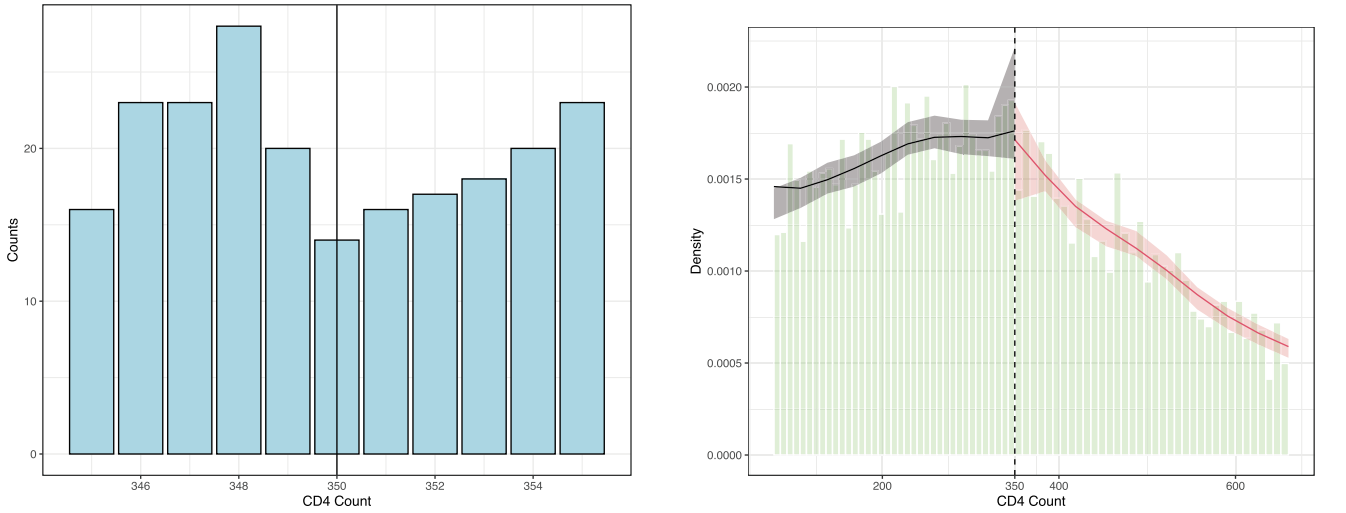
Note: Each row reports the average effect (at the cutoff) of being assigned to the treatment vs the control condition on a given predetermined covariate. Analysis based on local linear estimation with MSE-optimal bandwidth. The first and second columns report, respectively, the intercepts of the local linear fits to the left and right of the cutoff. The third column, $\widehat{\tau}_{\text{Y}}$, reports the difference between the first two columns, the intention-to-treat RD effect. *p*-value based on robust bias correction inference methods. The fifth column reports the MSE-optimal bandwidth. Column N_h^- reports the number of observations with score in $[c-h, c)$ and column N_h^+ reports the number of observations with score in $[c, c+h]$.

[330,370]; the results are already closer to results obtained with continuity-based methods in the smallest window, and nearly identical in the other two windows.

3.3 | Evaluating the RD assumptions

While one the strongest methods for causal inference and program evaluation, RD designs are ultimately a type of observational study and their key underlying assumptions are not guaranteed to hold by design.[53,54] The main threat to the validity of any RD design is the possibility of the units changing or "manipulating" their score in order to systematically select into the treatment. Analysts can offer supporting evidence in favor of the validity of the RD design in two main ways. First, investigators should provide qualitative information about the administrative process by which scores are assigned and cutoffs are determined—including whether this information is public knowledge. In medical applications, we might expect the RD design to be more robust when the score is a lab test. For example, in the ART example, patients might try to influence their CD4 count in order to qualify for ART. However, so long as the CD4 count is determined by laboratory procedures that cannot be precisely manipulated by patients or physicians, this is not a concern.

Second, the analysis of RD designs should include a series of falsification tests and diagnostics. As a general rule, falsification tests cannot prove that an assumption holds, but they can provide indirect empirical evidence that an assumption is likely to be invalid. Falsification tests arise from the fact that causal theories often predict an absence of treatment effects in addition to predicting the presence of such effects. We review several key falsification and diagnostic tests for RD designs, and illustrate their use with the ART application. All these methods are applicable to both sharp and fuzzy RD settings. In addition, similarly to IV settings, we stress the



(A) Histogram near cutoff: Score (X_i) within window [345, 355].

(B) Histogram and local polynomial density estimate for score (X_i) with 95% robust bias corrected confidence intervals.

FIGURE 4 Density plots for CD4 count (score) around the cutoff—ART application. The score X_i is patient i's CD4 count, and all patients below the cutoff are assigned to receive ART. In panel (B), the solid line is a local polynomial estimate of the density of X_i and the shaded regions represent 95% confidence intervals, both calculated separately for patients above and below the cutoff.

importance of checking the strength of the first-stage estimate in the fuzzy RD design (ie, τ_{D} and θ_{D} should be well-separated from zero). See Cattaneo et al[15,16] for more discussion, and Cattaneo et al[14] for an overview of the literature.

3.3.1 | Score density near the cutoff

This diagnostic test examines whether, in a local neighborhood near the cutoff, the number of observations below the cutoff is surprisingly different from the number of observations above it.[22] The underlying assumption is that if individuals do not have the ability to precisely manipulate the value of the score that they receive, the number of treated observations just above the cutoff should be approximately similar to the number of control observations below it. Although this assumption is neither necessary nor sufficient for the validity of an RD design, RD applications where there is an unexplained abrupt change in the number of observations at the cutoff will tend to be less credible.

This test is usually implemented in two ways, each motivated by one of the main two RD frameworks discussed in previous sections. The first method is the *Binomial Test* introduced by Cattaneo et al,[24,27] building on the local randomization framework. The second method is the *Density Test* introduced by McCrary,[22] which is based on the continuity-based framework for RD analysis. Informally, both tests seek to detect whether there is a significant amount of "bunching" at or near the cutoff. A small p-value under both tests indicates a significant amount of bunching, which is a concern. Cattaneo et al[55] develop a version of the density test based on local polynomial density estimation which can be plotted. All these tests are available in the software resources.

Figure 1A showed the raw histogram of the score in the ART application; in Figure 4A, we zoom in, showing the histogram only for the region [345,355]. Informally, there appear to be no obvious signs of bunching near the 350 cutoff. Formally, we do not reject the hypothesis of a change in density near the cutoff using the binomial test in the window [349,350] (p-value = 0.2026). We also implement the density test based on local polynomials, which we illustrate in Figure 4B. We fail to reject the null hypothesis that, at the cutoff, the limit of the score density from above the cutoff is the same as the limit from below ($p = 0.1858$). (For implementation, we used the `rddensity` and the `rdlocrand` software packages.) Overall, we find no evidence that the density of the score changes abruptly at or near the 350 cutoff and thus we see no evidence of intentional manipulation of the score.

3.3.2 | Predetermined covariates and placebo outcomes

Another important falsification test is based on the idea that if units lack the ability to precisely manipulate the value of their score, units just above and just below the cutoff should be similar in terms of all characteristics that could not have been affected by the treatment. These characteristics can be divided into two groups: *predetermined covariates*—variables that are determined before the treatment is assigned, and *placebo outcomes*—variables that are determined after the treatment is assigned but, according to substantive knowledge, could not have been affected by the treatment. In general, baseline covariates should be available in most applications, but the availability of placebo outcomes will vary from application to application.

This falsification test consists of repeating the RD analysis with baseline covariates or placebo outcomes in place of the main outcome of interest. The implementation can be done using both the continuity-based and local randomization frameworks. The underlying assumptions and methods for each case are analogous to those described previously, with the only change that now the outcome variable is either a predetermined covariate or a placebo outcome. As such, implementing this falsification test does not require any special software resources other than those for standard RD estimation and inference. With continuity-based methods, the implementation should use a bandwidth that is specific to each baseline covariate or placebo outcome, instead of the bandwidth selected for the main outcome of interest. In the local randomization framework, some predetermined covariates Z are used to select the window while others could be used for falsification testing after the local randomization window is selected; all falsification tests are conducted in the same chosen window. Regardless of the specific framework and methods employed, from the perspective of falsifying the RD design using predetermined covariates or placebo outcomes, the null hypothesis of no treatment effect should not be rejected in order to offer empirical evidence in favor of the RD assumptions.

We illustrate these ideas with the ART application, estimating the RD treatment effect for predetermined covariates. The results are based on the function `rdrobust` in the `rdrobust` package, which is used for estimation of RD effects and includes both bandwidth selection and robust bias correction inference methods. Table 2 analyzes the available baseline covariates in the data. The results are calculated using robust local polynomial methods to estimate RD effects, treating each predetermined covariate as an outcome. We find that covariate differences at the cutoff are generally quite small and none of the p-values are below 0.10. These results are reassuring, as they do not show signs of systematic differences near the cutoff: patients just above the cutoff are similar in terms of baseline covariates to patients just below the cutoff. Similar results are obtained when using local randomization methods, which we omit to conserve space.

3.3.3 | Bandwidth sensitivity, donut hole, and placebo cutoffs

This battery of diagnostic tests have all the same underling principle: they investigate the sensitivity of the results to small changes of different features of the implementations and data. The tests consider whether varying the bandwidth or local randomization neighborhood changes the empirical results; whether the observations closest to the cutoff overwhelmingly affect the extrapolation; and whether a fake treatment assignment rule (cutoff) leads to nonzero treatment effects.

The first method focuses on *bandwidth or local randomization neighborhood sensitivity*, which is a common strategy to probe the robustness of the empirical conclusions results to variations of the local neighborhood used for analysis. For example, as discussed previously for continuity-based methods, a larger bandwidth will on average lead to a more precise but also more biased RD treatment effect, if misspecification of the unknown conditional expectations approximations near the cutoff is a concern. To illustrate using the ART application, Table 3 reports results based on two wider bandwidths (relative to the MSE-optimal choice in Table 1); the results show that the conclusions are robust: treatment effects and their associated statistical significance remain qualitatively unchanged. A similar procedure can be used to probe the local randomization window, which can also be shrunk or enlarged to investigate the sensitivity of the main empirical findings; see Table 4.

A related falsification method is the so-called *donut hole sensitivity* method, which is based on the idea that the few observations closest to the cutoff should not drastically determine the empirical results. This is the mirror image of the bandwidth sensitivity: in both cases some observations are included or excluded depending on their score values relative to the cutoff. The donut hole falsification test removes a few observations closest to the cutoff in an attempt to

TABLE 3 Continuity-based sensitivity diagnostics—ART application.

	RD Effect	95% Robust CI	Bandwidth (h)	N_h^-	N_h^+
Bandwidth sensitivity check[a]					
ITT effect of ART assignment on ART initiation	−0.22	[−0.28, −0.13]	124.36	1634	1303
ITT effect of ART assignment on program retention	−0.15	[−0.22, −0.05]	124.36	1634	1303
Fuzzy effect of ART initiation on program retention	0.68	[0.37,1.00]	124.36	1634	1303
ITT effect of ART assignment on ART initiation	−0.23	[−0.29, −0.14]	149.36	1965	1525
ITT effect of ART assignment on program retention	−0.16	[−0.23, −0.07]	149.36	1965	1525
Fuzzy effect of ART initiation on program retention	0.69	[0.42,0.98]	149.36	1965	1525
Donut hole diagnostic[b]					
ITT effect of ART assignment on ART initiation	−0.23	[−0.3, −0.14]	122.24	1594	1262
ITT effect of ART assignment on program retention	−0.14	[−0.21, −0.04]	122.24	1594	1262
Fuzzy effect of ART initiation on program retention	0.59	[0.27,0.89]	122.24	1594	1262
Placebo cutoffs diagnostic $c = 300$					
ITT effect of ART assignment on ART initiation	−0.02	[−0.19,0.13]	29.95	551	547
ITT effect of ART assignment on program retention	0.03	[−0.14,0.21]	35.91	467	442
Placebo cutoffs diagnostic $c = 400$					
ITT effect of ART assignment on ART initiation	−0.01	[−0.10,0.05]	39.11	676	571
ITT effect of ART assignment on program retention	0.00	[−0.16,0.18]	44.13	528	413

Note: The two/three rows in each panel show, respectively, $\hat{\tau}_{\text{D}}$, $\hat{\tau}_{\text{Y}}$, and $\hat{\tau}_{\text{FRD}}$. Analysis based on local linear estimation with MSE-optimal main bandwidth reported in third column. Column labeled "95% Robust CI" reports the robust 95% confidence intervals based on robust bias-corrected inference. Column N_h^- reports the number of observations with score in $[c-h, c)$ and column N_h^+ reports the number of observations with score in $[c, c+h]$.

[a] Bandwidth used in the sensitivity check are ±10 relative to the benchmark MSE-optimal bandwidth reported in Table 1.

[b] Analysis based on local linear estimation with MSE-optimal main bandwidth reported in third column, but excluding observations with CD4 count equal to 349, 350, and 351.

understand the sensitivity of the results to those observations, since polynomial approximations can suffer from biases near the cutoff because of Runge's phenomenon. In practice, this method is easily implemented by using either the continuity-based framework or the local randomization framework, using different subsamples where observations in a symmetric interval around the cutoff are removed, starting with those closest to the cutoff and then progressing with larger intervals around cutoff. No special software is needed beyond the packages used for RD treatment effect estimation and inference (`rdrobust`, `rdlocrand`). Importantly, unlike the case of predetermined covariates and placebo outcomes, the same bandwidth or local randomization window used for treatment effect estimation should be used, instead of re-estimating a new bandwidth or window for each new subsample generated by the donut hole. We illustrate the donut hole diagnostic test with the ART application, dropping patients with CD4 count values of 349, 350, and 351 and re-estimating the RD effects using continuity-based methods only to conserve space. We report the results in Table 3, which show only minor differences between the donut hole estimates and the main results. This implies that our results are not sensitive to the small set of patients with CD4 counts right around the cutoff.

A third sensitivity approach investigates placebo cutoffs using either only control or only treated observations. The idea is to provide evidence in favor of continuity of the regression functions or, more generally, validity of the treatment assignment rule. In a nutshell, this approach analyzes either control or treatment units separately, and sets a sequence of artificial or placebo RD cutoffs to check that there is no RD treatment effect at those alternative cutoffs, since the expectation is that a treatment effect should occur only at the true cutoff and not at artificial cutoffs where treatment status is constant by construction. Empirical evidence of treatment effects at artificial cutoffs may undermine the design if the researcher cannot explain why these effects occur: nonzero effects at artificial cutoffs suggest the possibility that other factors are affecting the units in the background. Table 3 illustrates the idea with placebo cutoffs 300 and 400,

TABLE 4 Local randomization neighborhood sensitivity diagnostic—ART application.

	Risk difference	95% confidence interval	$N_{\mathcal{W}}^-$	$N_{\mathcal{W}}^+$
$\mathcal{W}$ = [340,360]				
ITT effect of ART assignment on ART initiation	−0.14	[−0.25, −0.04]	144	127
ITT effect of ART assignment on program retention	−0.09	[−0.20, 0.030]	144	127
Fuzzy effect of ART initiation on program retention	0.60	[−0.06, 1.27]	144	127
$\mathcal{W}$ = [335,365]				
ITT effect of ART assignment on ART initiation	−0.21	[−0.30, −0.13]	212	189
ITT effect of ART assignment on program retention	−0.11	[−0.20, −0.01]	212	189
Fuzzy effect of ART initiation on program retention	0.50	[0.13, 0.86]	212	189
$\mathcal{W}$ = [330,370]				
ITT effect of ART assignment on ART initiation	−0.25	[−0.32, −0.18]	277	245
ITT effect of ART assignment on program retention	−0.13	[−0.21, −0.06]	277	245
Fuzzy effect of ART initiation on program retention	0.52	[0.25, 0.79]	277	245

Note: The three rows in each panel show, respectively, $\widehat{\theta}_{\text{D}}$, $\widehat{\theta}_{\text{Y}}$, and $\widehat{\theta}_{\text{FRD}}$, corresponding to the local randomization estimates based on local randomization window $\mathcal{W}$. Benchmark local randomization window is reported in Table 1. Column $N_{\mathcal{W}}^-$ reports the number of observations with score in $\mathcal{W}$ and below the cutoff ($T_i = 0$) and column $N_{\mathcal{W}}^+$ reports the number of observations with score in $\mathcal{W}$ and above the cutoff ($T_i = 1$).

using continuity-based methods only to conserve space. For both intention-to-treat effects on program retention and ART initiation, we find that the robust 95% confidence intervals include zero, a reassuring result.

All the empirical results in this section are based on varying arguments in the `rdrobust` package for continuity-based methods, and in `rdlocrand` package for local randomization methods, and hence are readily available using general purpose software. See accompanying replication files.

3.3.4 | Fuzzy RD validation

To close our discussion of RD falsification methods, we review some validation methods that are specific to fuzzy RD designs. Since the fuzzy RD design shares several features with the IV design, these tests are generally based on diagnostics methods for IV designs. See References 31,56-58, and references therein, for reviews and examples of empirical evaluation of IV assumptions in biomedical research and causal inference.

The canonical fuzzy RD estimator is a local version of the standard two-stage least squares estimator in IV settings, and hence it requires a first stage well-separated from zero. Failure of this condition leads to a problem known as "weak instruments" in the IV literature, and is a serious concern when analyzing fuzzy RD designs.[59] In IV designs, a weak IV test is used to ensure that the effect of the instrument on the treatment exposure is sufficiently strong. The validation analysis of the fuzzy RD design should include this test, with the key difference that the standard weak IV test should be applied within the local neighborhood around the cutoff. Performing the test in a local neighborhood is important: a weak IV test that uses all observations is likely to overstate the strength of the instrument, since it would include data that is excluded from the main analysis by bandwidth or window selection.

Similarly to the sharp RD design, predetermined covariates should be used to validate the assumptions of the fuzzy RD design. These tests are implemented in the same way as in the sharp RD design, exploring whether the covariates are balanced in a neighborhood of the cutoff. These balance tests should use only the predetermined covariates and the score; the treatment received should not be used for covariate falsification purposes. Another IV diagnostic reports some measure of bias associated with the instrument rather than balance, since bias from a baseline covariate can be amplified by how strongly the IV is predictive of the treatment. This second diagnostic approach can be easily accommodated to the RD setting by applying the fuzzy RD ratio estimator to the baseline covariates (instead of the standard sharp RD estimator) when testing for differences in baseline covariates. See Davies et al[60] and Branson and Keele[61] for related formal and graphical methods, which we omit to conserve space.

TABLE 5 Continuity-based fuzzy RD estimates for predetermined covariates with robust bias corrected inference—ART application.

	$\hat{\tau}_{\text{FRD}}$	Robust *p*-value	MSE-optimal bandwidth	N_h^-	N_h^+
Age 0-18	−0.05	0.41	122.00	2281	1835
Age 18-25	−0.10	0.58	91.15	1718	1440
Age 25-30	0.14	0.37	111.16	2093	1674
Age 30-35	0.01	0.88	94.99	1767	1473
Age 35-40	−0.02	0.89	110.41	2073	1660
Age 40-45	−0.01	0.88	85.55	1592	1357
Age 45-55	0.00	0.90	105.77	1981	1606
Age 55+	0.01	0.83	114.10	2138	1725
2011 Qtr3	0.09	0.33	112.25	2152	1717
2011 Qtr4	0.09	0.54	79.37	1505	1299
2012 Qtr1	−0.09	0.45	106.25	2034	1646
2012 Qtr2	−0.00	0.99	85.82	1624	1383
2012 Qtr3	−0.05	0.64	114.33	2182	1755
2012 Qtr4	0.03	0.70	99.25	1891	1569
Female	−0.27	0.16	76.70	1443	1243
Clinic A	0.20	0.13	74.44	1416	1229
Clinic B	−0.11	0.38	85.94	1624	1383
Clinic C	−0.03	0.69	110.76	2116	1690

Note: The first column is the fuzzy RD effect ($\hat{\tau}_{\text{FRD}}$) for each predetermined covariate. Analysis based on local linear estimation with MSE-optimal bandwidth. *p*-value based on robust bias correction inference methods. The third column reports the MSE-optimal bandwidth. Column N_h^- reports the number of observations with score in $[c-h, c)$ and column N_h^+ reports the number of observations with score in $[c, c+h]$.

The key assumption known as the exclusion restriction—that being assigned to the treatment has no effect on the outcome except via actual treatment exposure—is untestable. This is a fundamental identifying assumption. For fuzzy RD designs, it is important to evaluate the exclusion restriction using qualitative information. See Arai et al[30] and references therein.

We illustrate these falsification methods with the ART application. We focus on two tests that are specific to RD designs with noncompliance: a weak instrument test, and covariate "balance" tests based on the ratio fuzzy RD estimator. We implement a weak IV test in the following way. First, we estimate the MSE-optimal bandwidth for both sides of the cutoff, using ART initiation (the treatment) as the outcome. We then implement a weak IV test using only the observations within this bandwidth. This test consists of regressing the treatment on an indicator variable for whether the CD4 score is less than 350, and assessing the *F*-value from this regression. The *F*-value from the weak IV test is approximately 698, well above the standard critical value thresholds used in the IV literature. Standard methods can be used to generate these *F*-values while using bandwidth selection methods in rdrobust; see the replication materials for details. We also test for differences in baseline covariates at the cutoff using the fuzzy RD estimator in Table 5. The balance results in Table 2 inflate the estimates in Table 2 by the strength of the instrument, which may make it more likely to find imbalances. However, the results are similar to those results in Table 2 based on the intention-to-treat RD estimator: we still find that there are no significant differences in the baseline covariates at the cutoff.

4 | ANALYSIS WITH DISCRETE SCORE

When the score only exhibits a "few" unique values (30 or fewer in typical applications), it is more appropriate to think of the RD design as truly having a discrete score. In that case, the methods presented in the previous sections need to be modified in order to be applicable. Continuity of the score is a key identifying assumption in the continuity-based

RD framework, since the continuity assumption is used to extrapolate the information of units near the cutoff under the thought experiment that, in large samples, eventually many units will be arbitrary close to the cutoff. Although continuity-based methods can be used under parametric assumptions when the score exhibits repeated values or mass points in its distribution, the local randomization RD framework is more naturally applicable in this setting because, by construction, this framework assumes valid extrapolation within the local randomization neighborhood—no further parametric assumptions are needed. Furthermore, the local randomization framework allows for employing only the closest observations to the cutoff, which tends to minimize extrapolation biases. In this section, we discuss how to employ both RD frameworks when the score is discrete with a few mass points, focusing only on the changes in interpretation and implementation that are required.

We consider a second biomedical empirical illustration on patient cost-sharing and healthcare utilization. In many countries, health care costs are subsidized through government programs, and research in health policy and management seeks to understand whether lower levels of cost-sharing encourages patients to use healthcare services at higher rates. Government health care cost-sharing often varies by age, thereby creating a discontinuity in health care subsidization that can be analyzed using RD methods. In this type of RD design, researchers compare health care utilization for those just above and below the age at which cost-sharing levels change. For example, in the U.S., eligibility for the healthcare program Medicare starts at age 65, and thus a common RD empirical strategy compares health care usage or other outcomes of interest for adults just above and below this age threshold. We reanalyze the recent work of Han et al,[62] who studied patient cost-sharing and healthcare utilization in Taiwan, where all inpatient and outpatient services for children under the age of 3 are exempt from co-payments. They used this discontinuity in age to compare levels of health care utilization just before and after the third birthday. Henceforth, we refer to this study as the *cost-sharing* application.

The raw data includes 414 282 children born between 2003 and 2004, and records the number of days until the child's third birthday—normalized to be zero on the day of the third birthday. Data on healthcare utilization was collected for up to 180 days before and after each child's third birthday. The treatment is an indicator equal to 1 if the child's age at the time of their visit is 3 or greater. This captures the higher level of patient's cost-sharing due to the expiration of the subsidy after the third birthday. The outcome of interest is the number of medical visits per 10 000 person days, and thus the outcome variable Y_i records the average for all children within a day relative to the cutoff. The data was aggregated at the day-level, but the score records only the associated week for each day, with a total of 50 weeks. While the unit of analysis is at the day-level, the score variable records the week relative to the cutoff, so there are seven observations for each value of the score $X_i \in \{-25, -24, \dots, 0, \dots, 23, 24\}$. Unlike the ART application, this RD design is sharp because once the child is 3 years of age or older there are no exceptions to the change in cost-sharing.

Graphical presentation continues to be important when the score is discrete. However, in this case, histograms, scatterplots, and RD plots can be constructed directly by employing the unique values of the score; there is no need for binning the score data first as in Section 3. In the case of the cost-sharing application, the histogram would not be informative since each mass point contains exactly seven observations. Thus, Figure 5 presents a scatter plot and a RD plot instead.

Figure 5A shows that indeed seven observations share each unique value of the score. Furthermore, the scatter plot suggests that when children are three or older, the rate of hospital visits drops. Figure 5B presents an RD plot, where each bin is equivalent to one value of the score, so each dot reports the sample average of the seven observations for that week. More generally, if the score variable takes on the values $\{\mathsf{x}_{K_-}, \dots, \mathsf{x}_{-2}, \mathsf{x}_{-1}, c, \mathsf{x}_1, \mathsf{x}_2, \dots, \mathsf{x}_{K_+}\}$, with K_- denoting the number of unique values below the cutoff and K_+ denoting the number of unique values above the cutoff, then the RD plot reports $K = K_- + K_+ + 1$ sample averages of the outcome for each score value, $\{\overline{Y}_{K_-}, \dots, \overline{Y}_{-2}, \overline{Y}_{-1}, \overline{Y}_c, \overline{Y}_1, \overline{Y}_2, \dots, \overline{Y}_{K_+}\}$, where $\overline{Y}_j = \frac{1}{\#\{i : x_i = \mathsf{x}_j\}} \sum_{i : x_i = \mathsf{x}_j} Y_i$ for $j \in \{K_-, \dots, -2, -1, c, 1, 2, \dots, K_+\}$ and $\#\{A\}$ denotes the number of elements in the set A. The global polynomial fits are added for visual presentation only, as they represent a global parametric interpolation across the K unique values or mass points available for the score.

4.1 | Continuity-based methods

In RD settings where there are repeated score values among the units, continuity-based methods can be applied if the score takes on a relatively large number of distinct values and the researcher is willing to make additional assumptions. A reasonable approach is to view the number of unique values of the score K as the effective sample size, and thus treat the units with identical scores as independent measurements of each particular score level. From this perspective, the total sample size continues to be n but the effective sample size is smaller because there is only $K \leq n$ unique values among $X_1, X_2, \dots, X_n$.

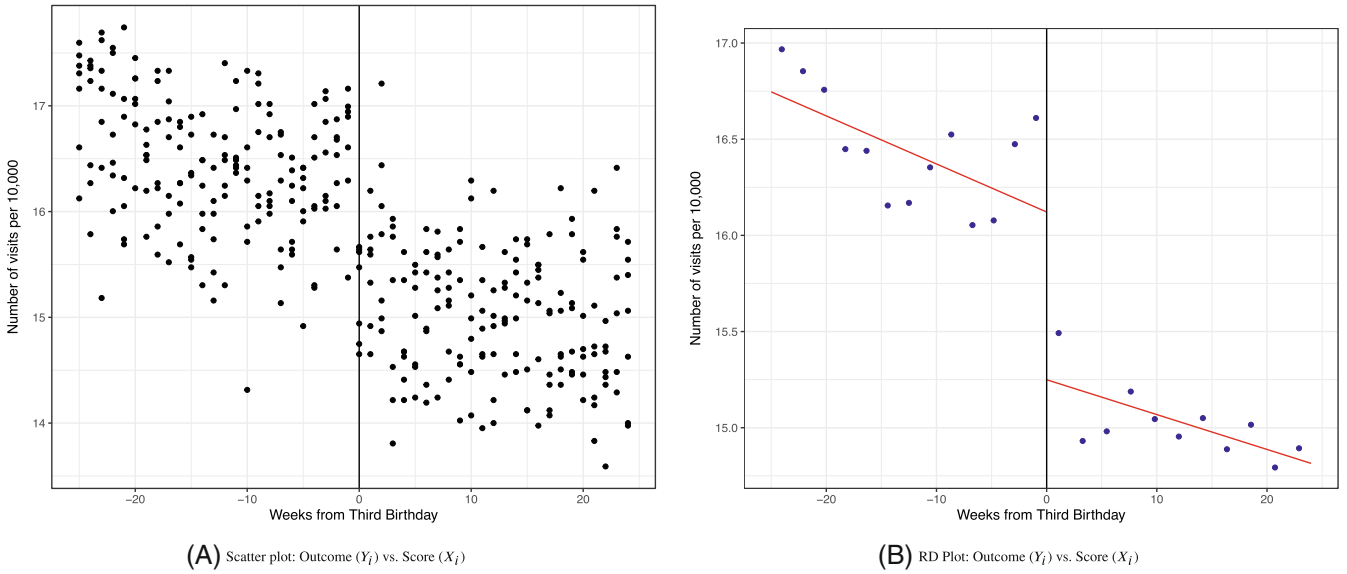


FIGURE 5 Basic plots—Cost-sharing application. The score X_i is the number of weeks until the child turns five years of age, where the date of birth is normalized to zero, so that positive numbers represents weeks past the third birthday, and negative numbers represent weeks before the third birthday. The cutoff is zero. Children below the cutoff are eligible for a health care subsidy; the subsidy is not available for children above the cutoff. The outcome Y_i is the number of medical visits per 10 000 person days. Both panels display the outcome against the score. In panel (B), dots are local means of Y_i calculated in different nonoverlapping bins of X_i, and the solid line is a 4th-order polynomial of Y_i on X_i, fitted separately for patients above and below the cutoff.

Whenever the number of the unique score values K is large enough and the closest mass points to the cutoff are close enough, the continuity-based framework can be taken as a reasonable approximation for identification purposes and hence τ_{SRD}, τ_{Y}, τ_{D}, and τ_{FRD} are reasonable treatment effects of interest. The key requirement is that whatever extrapolation takes place from the closest observed score value to the cutoff has a sufficiently small error. Clearly, some extrapolation would be needed, which is ultimately achieved via the estimation and inference methods employed.

Local polynomial methodology can be adapted and used for both optimal point estimation and robust bias corrected inference in settings with a large number of unique score values. The only important change is related to the sample size used, in addition to the key identifying assumptions invoked. From an approximation error perspective, only variation in the score variable can reveal the shape of the conditional expectation functions, and hence the correct sample size to be considered is K, not n. On the other hand, from an uncertainty perspective, either K or n could be the correct sample size, depending on the assumptions imposed about how the data was generated. Either way, the presence of repeated score values in the sample affects bandwidth selection and inference methods, but the necessary modifications are straightforward and readily applicable in general purpose software (`rdrobust` and `rddensity` packages). This logic justifies the empirical analysis in Section 3, where we used continuity-based methods despite having some repeated score values, because K was large. See Reference 16(section 3) and references therein.

The situation is different when K is small, that is, when there are only a few unique values in the score (usually 30 or less). In this case, it may be unreasonable to assume valid nonparametric extrapolation to the cutoff because it would be hard (or impossible) to learn the shape of the conditional expectation functions arbitrarily close to the cutoff. In this scenario, a solution to validate continuity-based methods is to rely on parametric extrapolation, where by virtue of the coarseness of the score, the postulated local polynomial model must be assumed to be correctly specified. This is a strong assumption, but necessary to restore point identification of RD treatment effect parameters at the cutoff.

To illustrate the point with an extreme example, suppose the score X_i takes only on five distinct values $\mathsf{x}_{-2} < \mathsf{x}_{-1} < c < \mathsf{x}_1 < \mathsf{x}_2$, where c continues to denote the RD cutoff. It follows that only $\mathbb{E}[Y_i(0)|X_i = \mathsf{x}_{-2}]$, $\mathbb{E}[Y_i(0)|X_i = \mathsf{x}_{-1}]$, $\mathbb{E}[Y_i(1)|X_i = c]$, $\mathbb{E}[Y_i(1)|X_i = \mathsf{x}_1]$ and $\mathbb{E}[Y_i(1)|X_i = \mathsf{x}_2]$ are identifiable from the data. In particular, $\tau_{\text{SRD}} = \mathbb{E}[Y_i(1)|X_i = c] - \mathbb{E}[Y_i(0)|X_i = c]$ will never be nonparametrically identifiable because $\mathbb{E}[Y_i(0)|X_i = c]$ is not identifiable without parametric assumptions about the functional form of $\mathbb{E}[Y_i(0)|X_i = x]$ for $x \in (\mathsf{x}_{-1}, c]$. Moreover, $\mathbb{E}[Y_i(1)|X_i = c]$ will be nonparametrically identifiable in a super-population sense only in settings where $\mathbb{P}[X_i = c] > 0$, that is, when the number of repeated values at $X_i = c$ is sufficiently large. This example shows a more general phenomenon: if the score is discrete, the canonical

TABLE 6 Continuity-based sharp RD methods—Cost-sharing application.

	RD effect	95% Robust CI	Bandwidth (h)	N_h^-	N_h^+
Number of doctor visits per 10 000	−1.29	[−2.08, −0.75]	6.08	42	49

Note: Analysis based on local linear estimation with MSE-optimal main bandwidth reported in third column. Column labeled "95% Robust CI" reports the robust 95% confidence intervals based on robust bias-corrected inference. Column N_h^- reports the number of observations with score in $[c-h, c)$ and column N_h^+ reports the number of observations with score in $[c, c+h]$.

continuity-based RD parameters τ_{SRD}, τ_{Y}, τ_{D}, or τ_{FRD} are not point identifiable without strong, parametric assumptions about the functional form of $\mathbb{E}[Y_i(0)|X_i = x]$ and $\mathbb{E}[Y_i(1)|X_i = x]$. This leads to two possible conceptual approaches: (i) assume such parametric assumptions hold, or (ii) change the parameter of interest. As already discussed, continuity-based RD methods can be deployed to RD designs with discrete score variables whenever the local parametrizations are assumed to generate small misspecification bias, that is, when the local polynomial model is assumed to be approximately correctly specified.

We illustrate these ideas with the cost-sharing application. The RD effect is −1.29 with robust confidence interval of [−2.08, −0.75] and robust p-value of zero (main MSE-optimal bandwidth equal to 6.08 weeks). See Table 6 for the full results.

4.2 | Local randomization methods

Provided the parameter of interest is changed or reinterpreted appropriately, RD identification, estimation and inference under a local randomization framework remains valid when the score exhibits mass points. To formalize the core ideas, we continue to assume that the support of the score variable is $\{\mathsf{x}_{K_-}, \ldots, \mathsf{x}_{-2}, \mathsf{x}_{-1}, \mathsf{x}_c, \mathsf{x}_1, \mathsf{x}_2, \ldots, \mathsf{x}_{K_+}\}$, with $K = K_- + K_+ + 1$ the total number of unique values. The local randomization assumption reduces to specifying a window containing some of these unique values where the two local randomization conditions discussed in Section 2 are assumed to hold.

We can define the following alternative RD parameters for settings where the score has few mass points: $\tilde{\theta}_{\text{SRD}} = \mathbb{E}[Y_i(1)|X_i = c] - \mathbb{E}[Y_i(0)|X_i = \mathsf{x}_{-1}]$, $\tilde{\theta}_{\text{Y}} = \mathbb{E}[Y_i(1, D_i(1))|X_i = c] - \mathbb{E}[Y_i(0, D_i(0))|X_i = \mathsf{x}_{-1}]$, $\tilde{\theta}_{\text{D}} = \mathbb{E}[D_i(1)|X_i = c] - \mathbb{E}[D_i(0)|X_i = \mathsf{x}_{-1}]$, and $\tilde{\theta}_{\text{FRD}} = \tilde{\theta}_{\text{Y}}/\tilde{\theta}_{\text{D}}$. The notation makes clear that the parameters of interest have changed: they now correspond to comparisons of potential outcomes at different values of the score variable ($X_i = c$ vs $X_i = \mathsf{x}_{-1}$). This approach allows for the deployment of local randomization RD methods. First, because the Fisherian approach is finite-sample valid, this method can be used even with small sample size at the two score evaluation points $X_i = c$ and $X_i = \mathsf{x}_{-1}$. The super-population approach, in contrast, relies on large sample approximations and consequently requires a large enough number of repeated values at $X_i = c$ and $X_i = \mathsf{x}_{-1}$. In practice, this idea can be used for the two closest values to the cutoff or, alternatively, for a collection of unique values closest to the cutoff. As before, the number of unique points on the score closest to the cutoff used is determined by the choice of window $\mathcal{W}$.

The choice of $\mathcal{W}$ in this case is simplified considerably. The implementation of the window selector based on covariates should start with the smallest possible window, $[\mathsf{x}_{-1}, c]$, and continue increasing this window one mass point at a time on either side. If there are enough observations in the window $[\mathsf{x}_{-1}, c]$, researchers should report results for this window. Even if a larger window is chosen by the covariate-based window selector, it will be important to show the results when only the observations closest to the cutoff are included in the analysis. Whenever $\mathcal{W}$ contains enough unique values of the score, it is also possible to use parametric extrapolation ideas. In this case, a parametric relationship is postulated between the outcome variables and the score, and regression-based methods are used for adjustment.

We illustrate these ideas with the cost-sharing application. We use the local randomization approach and implement a window selector to find the largest window around the cutoff where all covariates are balanced in that window and in all the windows contained in it. We use four predetermined covariates in our window selector: share of male children, household income per capita, share of children born in Taipei, and birth year. The results show that only the first window, which has seven observations on each side, has all covariates balanced; starting in the second window (14 days on either side of the cutoff), the minimum p-value is well under 5% (in fact, it is zero for all windows after the second). Table 7 shows the results of the balance tests in our selected window, where not only are the p-values above a 0.15 threshold, but the differences in means are very small for all four covariates.

We now estimate the effect of the treatment in the selected window. Since the number of observations in this window is only 14, it is important that we use Fisherian methods for inference, since those do not rely on large-sample

TABLE 7 Distribution of predetermined covariates for $\mathcal{W} = [-1, 0]$—Cost-sharing application.

	Mean below	Mean above	Diff. in means	*p*-value
Share of male	0.55	0.55	0.00	0.78
Household income per capita	12 494.53	12 532.86	38.33	0.21
Share of children born in Taipei	0.08	0.08	0.00	0.80
Birth year	2003.49	2003.49	−0.00	0.21

Note: Only including children who are within 7 days of their third birthday; there are 14 total observations, 7 on each side of the cutoff. The last column shows the Fisherian *p*-value assuming a fixed margins randomization mechanism that assigns these 14 observations to be above or below the birthday cutoff (which is normalized at zero).

TABLE 8 Local randomization sharp RD methods—Cost-sharing application.

Local randomization estimate				
	Mean below	**Mean above**	**Diff. in means**	**Fisherian *p*-value**
Number of doctor visits per 10 000	16.610	15.248	−1.362	0.006
Local randomization placebo estimate (pretreatment period)				
Number of doctor visits per 10 000	11.918	11.987	0.069	0.720

Note: Only including children who are within 7 days of their third birthday; there are 14 total observations, 7 on each side of the cutoff. The last column shows the Fisherian *p*-value assuming a fixed margins randomization mechanism that assigns these 14 observations to be above or below the birthday cutoff (which is normalized at zero).

approximations and provide exact *p*-values even when sample sizes are very small as in this case. The results are shown in Table 8, where we see that the mean difference in the number of doctor visits per 10 000 is −1.362: the number of visits per 10 000 is 16.61 for children who are two years old and whose third birthday is within 7 days, compared to 15.248 for children who turned three in the past seven days. The Fisherian *p*-value associated with the test of the hypothesis that there is no effect for any unit is well below 1%, despite the low sample size. This suggests, as expected, that cost sharing causes families to use medical care at higher rates.

Finally, we conduct a placebo analysis to assess the validity of the design. The data contains information on healthcare utilization for the period from 1997 to 2002, when cost-sharing was not higher for children under three. Thus, we expect no treatment effects from a similar analysis in this pretreatment period. Indeed, Table 8 shows that the Fisherian sharp null of no treatment effect cannot be rejected, with a *p*-value of 0.209. Moreover, the difference in means of −0.127 is less than one tenth of the −1.362 difference observed in the posttreatment period. Similar null results are obtained if the window in the placebo analysis is widened to plus or minus four weeks of the date of the third birthday.

4.3 | Evaluating the RD assumptions

In Section 3, we discussed an array of falsification and validation methods for RD designs with a continuous score variable. Those methods can be directly employed in settings where the score is discrete but the number of unique score values K is large enough. When K is small, some of these methods are easily applicable while others are not. For the *score density near the cutoff* diagnostic, the binomial test continues to be valid regardless of the size of K because this is a finite-sample valid test about the relative proportion of units on either side of the cutoff. On the other hand, the density test must be handled with more care because that method was developed for (approximately) continuously distributed scores. For *predetermined covariates and placebo outcomes* diagnostics, all ideas and methods discussed in this section can be applied directly. *Bandwidth sensitivity* diagnostics can be applied when the score is discrete, while the *donut hole* and *placebo cutoff* diagnostic are more difficult to implement without strong parametric assumptions. Finally, *fuzzy RD validation* diagnostics can be adapted from the standard IV literature straightforwardly.

4.4 | A flawed RD application: Genetic assay guidelines for chemotherapy

The ART and cost-sharing applications showcase RD designs with many and with few discrete score values that pass all the key diagnostic tests, and produce robust and statistically significant treatment effect estimates. To contrast, we now discuss a third empirical application where the key falsification methods do not support the use of RD methods.

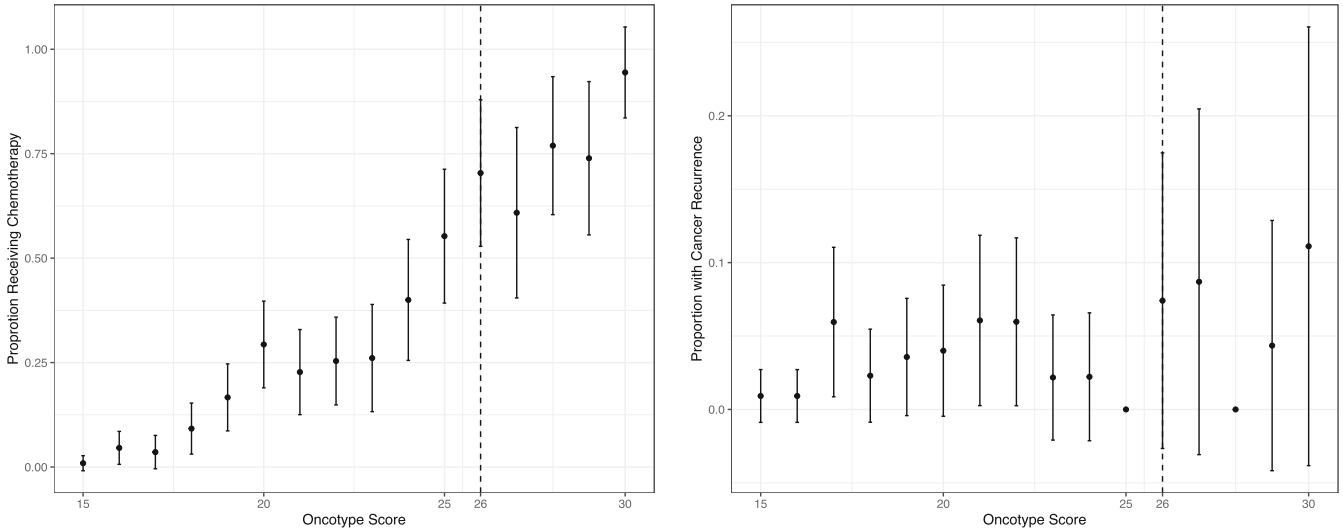
While treatment options for breast cancer have greatly expanded over the last two decades, chemotherapy is still often indicated for patients. To guide whether chemotherapy should be administered there are several commercially available gene-expression assays that provide prognostic information in hormone-receptor positive breast cancer patients. One widely used score is the Oncotype DX by Genomic Health, which is a 21-gene recurrence-score assay that ranges from 0 to 100 and is predictive of chemotherapy benefit when it is high—with a high score defined as 31 or higher. When the oncotype score is low (0 to 10), it is prognostic for a very low rate of distant breast cancer recurrence (2%) and adjuvant chemotherapy is not recommended. There is, however, considerable uncertainty as to whether chemotherapy is beneficial for patients who have a mid-range oncotype score. Current clinical guidelines suggest initiation of adjuvant chemotherapy for patients with an oncotype score of 26 or higher.[63-65] Thus, this setup suggests the use of an RD design with a discrete running variable being the oncotype score and the cutoff is 26.

For this application, we analyze a cohort of patients from the Penn Breast Database from 2009 to 2017 with oncotype scores of less than 40 who underwent surgery and were then eligible for adjuvant chemotherapy. Excluding patients with oncotype scores of 40 or greater reduces the cohort from 16 488 to 3269, after also excluding three patients who did not undergo oncotype scoring. The database includes several predetermined covariates: age, race, tumor size, tumor grade, an indicator for lymphovascular invasions, an indicator for estrogen receptor, an indicator for progesterone receptor, type of surgery (mastectomy or breast conservation), and an indicator for endocrine therapy. This is an RD design where the unit of observation is the patient, X_i is the oncotype score, $c = 26$, the treatment is the receipt of adjuvant chemotherapy, and the outcome of interest is an indicator for recurrence of breast cancer. This RD design is fuzzy since adjuvant chemotherapy was only prescribed to patients. Henceforth, we refer to this empirical application as the *chemotherapy* application.

Figure 6 illustrates the fuzzy RD design. Figure 6A shows the plot of the breast cancer recurrence indicator on the oncotype score, while Figure 6B shows the treatment take-up. Since this score takes only a smaller number of values, this is an example of an RD design with discrete score. We thus simply plot the proportion of patients with breast cancer recurrence for each one of these values. There is an increase in the proportion of patients that receive chemotherapy at the 26 cutoff, but not all patients with oncotype score of 26 receive it, and a considerable share of patients with oncotype scores below 26 are treated with chemotherapy. Moreover, the proportion of treated patients "jumps" not only at 26, but also at 25 and at 24, suggesting that some physicians are using a cutoff that is lower than the guideline, or perhaps not using a cutoff at all and simply steadily increasing the probability of chemotherapy treatment as the oncotype score increases. The evidence thus shows that many physicians initiate treatment for patients that are below the clinical guideline that is the basis for the RD design. In general, this pattern will tend to occur in applications where the physician deems the side effects of treatment to be small but the effects of treatment worthwhile. As we demonstrate below, this phenomenon will tend to make the instrument weak in the fuzzy RD design, and preclude the researcher's ability to learn about the treatment effect.

We attempt to validate the RD design by analyzing whether the number of treated and control observations is similar in a small neighborhood of the cutoff, by studying whether these observations are similar in terms of predetermined characteristics or covariates, and by investigating the presence of weak instruments. We implement the density test by applying the binomial test to the oncotype data. For the neighborhood of $W = [25, 26]$, there are 38 observations below the threshold and 27 above the threshold; assuming the null probability is 1/2, the p-value from the binomial test is 0.215. If instead we use the neighborhood $W = [24, 27]$, there are 83 observations with 50 above the threshold for a p-value of less than 0.005. The latter result is not consistent with a Bernoulli trial with probability of success 1/2. As oncotype scores increase, they become less common, and as such there is a clear downward trend in the density of the score. This results in a statistically significant imbalance in the number of units below and above the cutoff as soon as the second largest window is considered.

Next, we explore whether there is a window around the cutoff where all the covariates are balanced. We summarize the results without reporting details to conserve space. Considering the p-value for differences-in-means obtained for each increasing symmetric window from [25, 26] to [21, 30], we find that for the smallest window the p-value for one covariate, tumor size, is less than 0.05. If tumor size affects cancer recurrence, an imbalance in this covariate can invalidate the outcome comparisons between the treated and the control groups. Ideally, no important confounders should be imbalanced in the chosen local randomization window. Furthermore, the minimum p-value is below 0.10 in all of these



(A) RD Plot: Outcome (Y_i) vs. Score (X_i) with 95% asymptotic confidence intervals

(B) RD Plot: Treatment Take-up (D_i) vs. Score (X_i) with 95% asymptotic confidence intervals

FIGURE 6 Basic plots—Chemotherapy application. The score X_i is the Oncotype DX score for patient i. The cutoff is 26: the guideline is to initiate adjuvant chemotherapy for patients with a score of 26 or higher. The outcome Y_i is an indicator for recurrence of breast cancer. Panel (A) plots the proportion of patients receiving treatment against the score; panel (B) reports the outcome against the score. In both panels, the dots are sample means and the bars represent 95% confidence intervals.

windows, showing that covariate balance is getting worse as the window size increases—a pattern that is expected when the score correlates strongly with units' characteristics. We also found a second covariate, lymphovascular invasion, that is imbalanced in the second smallest window. These preintervention covariates are likely to be important determinants of the outcome, and thus the outcome of the RD methods are likely to be confounded and fail to provide a valid estimate of the true effect of chemotherapy on cancer recurrence. In sum, we find statistically significant differences in key preintervention features that are likely to confound the RD design.

Finally, we consider the first-stage treatment effect, θ_{D}, to understand whether reaching an oncotype score of 26 resulted in a significant increase in the probability of being treated with adjuvant chemotherapy. For the smallest window, $\mathcal{W} = [25, 26]$ with $N_{\mathcal{W}} = 65$, we find $\widehat{\theta}_{\mathrm{D}} = 0.15$ with a Fisherian p-value of 0.32. In addition, the large-sample F-statistic is 1.51, suggesting a clear problem of weak instruments, as already anticipated in Figure 6A. Reaching the cutoff of 26 does not seem to have induced an increase in the probability of receiving adjuvant chemotherapy. Fuzzy RD treatment effects are weakly identifiable and thus unreliable in this application.

In sum, the *chemotherapy* application does not pass basic RD validation/diagnostic tests, and the evidence does not support an RD analysis. The clinical guideline was not followed closely, which resulted in a very weak instrument, and the fact that important confounders are imbalanced even in the smallest windows around the cutoff makes this application an example of a flawed RD design.

5 | CONCLUSION

The RD design offers biomedical researchers the possibility of rigorously studying the effect of a treatment that is assigned based on a score and a cutoff, such as a recommendation to treat patients with a diagnostic laboratory test above or below a given threshold. Although the patient population above the cutoff will typically be very different from the patient population below the cutoff, the RD design restores comparability by focusing on patients whose scores are close to the cutoff on either side.

Our discussion, empirical examples, and accompanying computer code provides a state-of-the-art introduction and practical guide for the analysis of canonical RD designs in biomedical contexts. We covered modern estimation, inference, validation, and visualization approaches for the analysis and interpretation of RD designs based on both continuity-based and local randomization frameworks. The main takeaways are as follows: (i) always employ graphical approaches to

uncover basic features of the design but never for formal estimation and inference; (ii) always localize near the cutoff when deploying principled statistical methods and never rely on global estimation and inference approaches; (iii) depending on the coarseness of the score, employ continuity-based or local randomization methods being cognizant of the unavoidable extrapolation (to the cutoff) underlying the methods (eg, ART and Cost-Sharing applications); (iv) always employ falsification and validation diagnostics to offer evidence in favor of the RD design; and (v) be aware that not every treatment assignment mechanism based on a hard-thresholding rule can be analyzed via RD designs methods because sometimes the key underlying identifying assumptions do not hold (eg, Chemotherapy application).

The canonical RD design can be generalized to the case of multiple scores,[66,67] the geographic RD design,[68,69] multiple cutoffs,[26,29] kink RD designs,[70] RD designs with rounded scores,[71] and RD designs with measurement error,[72] among many other possibilities. We do not discuss all these extensions and generalization due to space constraints. See Cattaneo et al[14] for more examples and references, and Cattaneo et al[15,16] for practical introductions.

All the methods discussed in this tutorial employ open source general-purpose software for `R`, `Stata`, and `Python`, available at https://rdpackages.github.io/. We also provide full replication materials (data and codes) for the three applications, available at https://rdpackages.github.io/replication/.

ACKNOWLEDGEMENTS

We thank our current and former collaborators Sebastian Calonico, Max Farrell, Yingjie Feng, Brigham Frandsen, Nicolas Idrobo, Michael Jansson, Xinwei Ma, Kenichi Nagasawa, Filippo Palomba, Jasjeet Sekhon, and Gonzalo Vazquez-Bare for their intellectual input to our research program on RD designs. Cattaneo and Titiunik gratefully acknowledge financial support from the National Science Foundation (SES-2019432 and SES-2241575), and Cattaneo gratefully acknowledges financial support from the National Institute of Health (R01 GM072611-16).

DATA AVAILABILITY STATEMENT

We provide full replication materials at https://rdpackages.github.io/replication.

ORCID

Matias D. Cattaneo https://orcid.org/0000-0003-0493-7506
Luke Keele https://orcid.org/0000-0002-3859-2713
Rocío Titiunik https://orcid.org/0000-0001-5145-3059

REFERENCES

1. Craig P, Katikireddi SV, Leyland A, Popham F. Natural experiments: An overview of methods, approaches, and contributions to public health intervention research. *Annu Rev Public Health*. 2017;38:39-56.
2. Hernán MA. The C-word: Scientific euphemisms do not improve causal inference from observational data. *Am J Public Health*. 2018;108(5):616-619.
3. Hernán MA, Robins JM. *Causal Inference: What If*. Boca Raton: Chapman & Hall/CRC; 2022.
4. Rosenbaum PR. *Design of Observational Studies*. New York, NY: Springer; 2010.
5. Imbens GW, Rubin DB. *Causal Inference in Statistics, Social, and Biomedical Sciences*. New York, NY: Cambridge University Press; 2015.
6. Abadie A, Cattaneo MD. Econometric methods for program evaluation. *Annu Rev Econom*. 2018;10:465-503.
7. Bor J, Moscoe E, Mutevedzi P, Newell ML, Bärnighausen T. Regression discontinuity designs in epidemiology: Causal inference without randomized trials. *Epidemiology*. 2014;25(5):729-737.
8. O'Keeffe AG, Geneletti S, Baio G, Sharples LD, Nazareth I, Petersen I. Regression discontinuity designs: An approach to the evaluation of treatment efficacy in primary care using observational data. *BMJ*. 2014;349:g5293.
9. Bor J, Moscoe E, Bärnighausen T. Three approaches to causal inference in regression discontinuity designs. *Epidemiology*. 2015;26(2):e28-e30.
10. Maciejewski ML, Basu A. Regression discontinuity design. *JAMA*. 2020;324(4):381-382.
11. Titiunik R. Natural experiments. In: Druckman JN, Gree DP, eds. *Advances in Experimental Political Science*. New York, NY: Cambridge University Press; 2021:103-129.
12. Boon MH, Craig P, Thomson H, Campbell M, Moore L. Regression discontinuity designs in health: A systematic review. *Epidemiology*. 2021;32(1):87.
13. Thistlethwaite DL, Campbell DT. Regression-discontinuity analysis: An alternative to the ex-post facto experiment. *J Educ Psychol*. 1960;51(6):309-317.
14. Cattaneo MD, Titiunik R. Regression discontinuity designs. *Annu Rev Econom*. 2022;14:821-851.
15. Cattaneo MD, Idrobo N, Titiunik R. A practical introduction to regression discontinuity designs: Foundations. *Cambridge elements: Quantitative and Computational Methods for Social Science*. New York, NY: Cambridge University Press; 2020.

16. Cattaneo MD, Idrobo N, Titiunik R. A practical introduction to regression discontinuity designs: Extensions. *Cambridge Elements: Quantitative and Computational Methods for Social Science*. New York, NY: Cambridge University Press; 2023.
17. Bor J, Fox MP, Rosen S, et al. Treatment eligibility and retention in clinical HIV care: A regression discontinuity study in South Africa. *PLoS Med*. 2017;14(11):e1002463.
18. Tanser F, Hosegood V, Bärnighausen T, et al. Cohort profile: Africa centre demographic information system (ACDIS) and population-based HIV survey. *Int J Epidemiol*. 2007;37(5):956-962.
19. Houlihan CF, Bland RM, Mutevedzi PC, et al. Cohort profile: Hlabisa HIV treatment and care programme. *Int J Epidemiol*. 2010;40(2):318-326.
20. Calonico S, Cattaneo MD, Titiunik R. Optimal data-driven regression discontinuity plots. *J Am Stat Assoc*. 2015;110(512):1753-1769.
21. Korting C, Lieberman C, Matsudaira J, Pei Z, Shen Y. Visual inference and graphical representation in regression discontinuity designs. *Q J Econ*. 2023;138(3):1977-2019.
22. McCrary J. Manipulation of the running variable in the regression discontinuity design: A density test. *J Econ*. 2008;142(2):698-714.
23. Cattaneo MD, Crump RK, Farrell MH, Feng Y. On Binscatter. arXiv preprint, arXiv:1902.09608 2022.
24. Cattaneo MD, Titiunik R, Vazquez-Bare G. Comparing inference approaches for RD designs: A reexamination of the effect of head start on child mortality. *J Policy Anal Manage*. 2017;36(3):643-681.
25. Hahn J, Todd P, Klaauw vdW. Identification and estimation of treatment effects with a regression-discontinuity design. *Econometrica*. 2001;69(1):201-209.
26. Cattaneo MD, Keele L, Titiunik R, Vazquez-Bare G. Extrapolating treatment effects in multi-cutoff regression discontinuity designs. *J Am Stat Assoc*. 2021;116(536):1941-1952.
27. Cattaneo MD, Frandsen B, Titiunik R. Randomization inference in the regression discontinuity design: An application to party advantages in the U.S Senate. *J Causal Inference*. 2015;3(1):1-24.
28. Dong Y. Alternative assumptions to identify LATE in fuzzy regression discontinuity designs. *Oxf Bull Econ Stat*. 2018;80(5):1020-1027.
29. Cattaneo MD, Keele L, Titiunik R, Vazquez-Bare G. Interpreting regression discontinuity designs with multiple cutoffs. *J Polit*. 2016;78(4):1229-1248.
30. Arai Y, Hsu Y, Kitagawa T, Mourifié I, Wan Y. Testing identifying assumptions in fuzzy regression discontinuity designs. *Quant Econ*. 2022;13(1):1-28.
31. Baiocchi M, Cheng J, Small DS. Instrumental variable methods for causal inference. *Stat Med*. 2014;33(13):2297-2340.
32. Kaptchuk TJ, Miller FG. Placebo effects in medicine. *N Engl J Med*. 2015;373(1):8-9.
33. Fan J, Gijbels I. *Local polynomial Modelling and Its Applications*. Vol 66. Boca Raton, FL: CRC Press; 1996.
34. Calonico S, Cattaneo MD, Titiunik R. Robust nonparametric confidence intervals for regression-discontinuity designs. *Econometrica*. 2014;82(6):2295-2326.
35. Calonico S, Cattaneo MD, Farrell MH, Titiunik R. Regression discontinuity designs using covariates. *Rev Econ Stat*. 2019;101(3):442-451.
36. Cattaneo MD, Vazquez-Bare G. The Choice of neighborhood in regression discontinuity designs. *Observat Stud*. 2016;2:134-146.
37. Calonico S, Cattaneo MD, Farrell MH. Optimal bandwidth choice for robust bias corrected inference in regression discontinuity designs. *Econ J*. 2020;23(2):192-210.
38. Xu KL. Regression discontinuity with categorical outcomes. *J Econ*. 2017;201(1):1-18.
39. Arai Y, Ichimura H. Simultaneous selection of optimal bandwidths for the sharp regression discontinuity estimator. *Quant Econ*. 2018;9(1):441-482.
40. Dong Y, Lee YY, Gou M. Regression discontinuity designs with a continuous treatment. *J Am Stat Assoc*. 2021;118(541):208-221.
41. Arai Y, Otsu T, Seo MH. Regression discontinuity design with potentially many covariates. arXiv preprint, arXiv:2109.08351 2021.
42. Calonico S, Cattaneo MD, Farrell MH. On the effect of bias estimation on coverage accuracy in nonparametric inference. *J Am Stat Assoc*. 2018;113(522):767-779.
43. Calonico S, Cattaneo MD, Farrell MH. Coverage error optimal confidence intervals for local polynomial regression. *Bernoulli*. 2022;28(4):2998-3022.
44. Kamat V. On Nonparametric Inference in the Regression Discontinuity Design. *Economet Theor*. 2018;34(3):694-703.
45. Tuvaandorj P. Regression discontinuity designs, white noise models, and minimax. *J Econ*. 2020;218(2):587-608.
46. Ganong P, Jäger S. A permutation test for the regression kink design. *J Am Stat Assoc*. 2018;113(522):494-504.
47. Hyytinen A, Meriläinen J, Saarimaa T, Toivanen O, Tukiainen J. When does regression discontinuity design work? Evidence from random election outcomes. *Quant Econ*. 2018;9(2):1019-1051.
48. De Magalhães L, Hangartner D, Hirvonen S, Meriläinen J, Ruiz N, Tukiainen J. How much should we trust regression discontinuity design estimates? *Evidence from Experimental Benchmarks of the Incumbency Advantage*. working paper. Turku: Aboa Centre for Economics (ACE); 2020.
49. Cattaneo MD, Keele L, Titiunik R. Covariate adjustment in regression discontinuity designs. In: Small DS, Zubizarreta JR, Rosenbaum PR, eds. *Handbook of Matching and Weighting in Causal Inference*. Boca Raton, FL: Chapman & Hall; 2023.
50. Ernst MD. Permutation methods: A basis for exact inference. *Stat Sci*. 2004;19(4):676-685.
51. Keele LJ, Small DS, Grieve R. Randomization based instrumental variables methods for binary outcomes with an application to the IMPROVE trial. *J R Stat Soc Ser A*. 2017;180(2):569-586.
52. Kang H, Peck L, Keele L. Inference for instrumental variables: A randomization inference approach. *J R Stat Soc Ser A*. 2018;181(4):1231-1254.
53. Sekhon JS, Titiunik R. Understanding regression discontinuity designs as observational studies. *Observat Stud*. 2016;2:174-182.

54. Sekhon JS, Titiunik R. On interpreting the regression discontinuity design as a local experiment. In: Cattaneo MD, Escanciano JC, eds. *Regression Discontinuity Designs: Theory and Applications (Advances in Econometrics)*. Vol 38. Bingley, UK: Emerald Group Publishing; 2017:1-28.
55. Cattaneo MD, Jansson M, Ma X. Simple local polynomial density estimators. *J Am Stat Assoc*. 2020;115(531):1449-1455.
56. Glymour MM, Tchetgen Tchetgen EJ, Robins JM. Credible mendelian randomization studies: Approaches for evaluating the instrumental variable assumptions. *Am J Epidemiol*. 2012;175(4):332-339.
57. Pizer SD. Falsification testing of instrumental variables methods for comparative effectiveness research. *Health Serv Res*. 2016;51(2):790-811.
58. Keele LJ, Zhao Q, Kelz RR, Small DS. Falsification tests for instrumental variable designs with an application to tendency to operate. *Med Care*. 2019;57(2):167-171.
59. Feir D, Lemieux T, Marmer V. Weak identification in fuzzy regression discontinuity designs. *J Bus Econ Stat*. 2016;34(2):185-196.
60. Davies NM, Thomas KH, Taylor AE, et al. How to compare instrumental variable and conventional regression analyses using negative controls and bias plots. *Int J Epidemiol*. 2017;46(6):2067-2077.
61. Branson Z, Keele L. Evaluating a key instrumental variable assumption using randomization tests. *Am J Epidemiol*. 2020;189(11):1412-1420.
62. Han HW, Lien HM, Yang TT. Patient cost-sharing and healthcare utilization in early childhood: evidence from a regression discontinuity design. *Am Econ J Econ Pol*. 2020;12(3):238-278.
63. Paik S, Tang G, Shak S, et al. Gene expression and benefit of chemotherapy in women with node-negative, estrogen receptor–positive breast cancer. *J Clin Oncol*. 2006;24(23):3726-3734.
64. Sparano JA, Paik S. Development of the 21-gene assay and its application in clinical practice and clinical trials. *J Clin Oncol*. 2008;26(5):721-728.
65. Albain KS, Barlow WE, Shak S, et al. Prognostic and predictive value of the 21-gene recurrence score assay in postmenopausal women with node-positive, oestrogen-receptor-positive breast cancer on chemotherapy: A retrospective analysis of a randomised trial. *Lancet Oncol*. 2010;11(1):55-65.
66. Papay JP, Willett JB, Murnane RJ. Extending the regression-discontinuity approach to multiple assignment variables. *J Econ*. 2011;161(2):203-207.
67. Reardon SF, Robinson JP. Regression discontinuity designs with multiple rating-score variables. *J Res Educ Effect*. 2012;5(1):83-104.
68. Keele LJ, Titiunik R. Geographic boundaries as regression discontinuities. *Polit Anal*. 2015;23(1):127-155.
69. Keele LJ, Titiunik R, Zubizarreta J. Enhancing a geographic regression discontinuity design through matching to estimate the effect of ballot initiatives on voter turnout. *J R Stat Soc Ser A*. 2015;178(1):223-239.
70. Card D, Lee DS, Pei Z, Weber A. Inference on causal effects in a generalized regression kink design. *Econometrica*. 2015;83(6):2453-2483.
71. Barreca AI, Lindo JM, Waddell GR. Heaping-induced bias in regression-discontinuity designs. *Econ Inq*. 2016;54(1):268-293.
72. Bartalotti O, Brummet Q, Dieterle S. A correction for regression discontinuity designs with group-specific mismeasurement of the running variable. *J Bus Econ Stat*. 2021;39(3):833-848.

How to cite this article: Cattaneo MD, Keele L, Titiunik R. A guide to regression discontinuity designs in medical applications. *Statistics in Medicine*. 2023;1-30. doi: 10.1002/sim.9861

APPENDIX A. ANNOTATED SAMPLE CODE

Full replication files can be found at https://rdpackages.github.io/. The replication files include the data and scripts for `Python`, `R` and `Stata` to generate every result in the paper. Below we provide an abbreviated version of `R` code for the ART application for illustration purposes. Note that these commands will not exactly replicate the results in the tables due to handling of missing values. For exact replication of tables, see the replication codes.

```
> library(foreign)
> library(rdrobust)
> library(rddensity)
> library(rdlocrand)
>
> # Read dataset
> data <- read.dta("CKT_2023_SIM--ART.dta")
> X = data$cd4
> Y = data$visit_test_6_18
> D = data$art_6m
```

```
> ## Generate RD Plot
> rdplot(Y, X, c=350, y.label="Retained", x.label="CD4 Count", p=3)

> ## Continuity-Based RD Analysis: ITT effect on outcome
> rdrobust(Y, X, c=350)
> ## Continuity-Based RD Analysis: ITT effect on treatment
> rdrobust(D, X, c=350)
> ## Continuity-Based RD Analysis: Fuzzy effect on outcome
> rdrobust(Y, X, c=350, fuzzy=D)

> ## Local Randomization Window selection
> vars <- c("age1", "age2", "age3", "age4","age5", "age6", "age7", "age8", "qtr1",
  "qtr2",
+ "qtr3", "qtr4", "qtr5", "qtr6", "clinic_a", "clinic_b", "clinic_c")
> Z <- data[vars]
> out = rdwinselect(X, Z, c=350, seed = 50, reps = 1000, wstep=1)

> ## Local Randomization Analysis: ITT effect on outcome
> ci_vec = c(0.05, seq(from = -.5, to = .5, by = 0.01))
> rdrandinf(Y, X, cutoff = 350, seed= 5023, wl=346, wr=354, ci=ci_vec)

> ## Local Randomization Analysis: Fuzzy effect on outcome
> ci_vec = c(0.05, seq(from = -1, to = 1, by = 0.01))
> rdrandinf(Y, X, fuzzy=c(fuzzy.tr=D, fuzzy.stat="tsls"),
+ cutoff = 350, seed= 5023, wl=346, wr=354, ci=ci_vec)
```