A Guide to Regression Discontinuity Designs in Medical Applications *

Matias D. Cattaneo† Luke Keele‡ Rocío Titiunik§

May 24, 2022

Abstract

We present a guide to practice for the analysis of regression discontinuity (RD) designs in biomedical contexts, discussing modern validation, estimation, and inference methods based on both continuity and local randomization approaches. In addition to offering an introduction to the state-of-the-art in RD designs methods, we focus on two particular features that are relevant in biomedical research. First, we emphasize methods for fuzzy RD designs, which arise most often when therapeutic treatments are based on clinical guidelines. Second, we discuss how to analyze RD designs with discrete scores, which are ubiquitous in biomedical applications. We illustrate our discussion with three empirical examples: the effect CD4 guidelines for antiretroviral 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.

Keywords: treatment effect and policy evaluation, causal inference, regression discontinuity.

*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 Cattaneo gratefully acknowledges financial support from the National Institute of Health (R01 GM072611-16).

†Department of Operations Research and Financial Engineering, Princeton University.

‡Department of Surgery and Biostatistics, University of Pennsylvania.

§Department of Politics, Princeton University.

Contents

1 Introduction

Drawing causal inferences from quantitative data is a fundamental goal in epidemiology, comparative effectiveness, health services, and outcomes research (Craig et al., 2017; Hernán, 2018; Hernán and Robins, 2022). 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 (Rosenbaum, 2010; Imbens and Rubin, 2015; Abadie and Cattaneo, 2018; Hernán and Robins, 2022). Although the term *natural experiment* is defined differently across disciplines (Titiunik, 2021), all definitions involve some external circumstances that produce a seemingly arbitrary assignment of an intervention that allows comparing control and treated units to study treatment effects.

Over the last two decades, many natural experiments in biomedical research have been based on instrumental variables (Swanson and Hernán, 2013; Garabedian et al., 2014). Given the strong assumptions associated with instrumental variable designs, some scholars have recently advocated for greater use of a type of natural experiment known as the regression discontinuity (RD) design in biomedical contexts (Bor et al., 2014; O'Keeffe et al., 2014; Bor et al., 2015; Maciejewski and Basu, 2020). As a result, RD designs have become much more common in biomedical research: a recent review identified over 325 studies based on RD designs in medical studies (Boon et al., 2021).

The popularity of the RD designs stems from their high internal validity. Causal inferences from RD designs are often more credible and robust than those from other non-experimental impact evaluation strategies such as differences-in-differences or instrumental variables 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 and all units with scores below the cutoff are assigned to the control condition. In the

so-called *sharp* RD design, all units comply perfectly with the treatment 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.

The RD design was first introduced by Thistlethwaite and Campbell (1960) in education research to study the effect of a scholarship based on test scores. In biomedical contexts, RD designs commonly arise from treatment guidelines based on diagnostic test results. For instance, a specific treatment is recommended (or administered) when test results exceed a known cutoff—e.g. 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 of systematic differences in their 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.

The observation that, near the cutoff, chance should play a role in the units' placement above or below the cutoff was originally made by Thistlethwaite and Campbell (1960), and has since led to a strand of the literature that views RD designs as "local" randomized experiments (Lee, 2008; Cattaneo et al., 2015, 2017). This approach, which is usually referred to as the *local randomization framework* for RD designs, analyzes and interprets these designs using tools from the classical literature on the analysis of experiments. In contrast, a different approach commonly known as the *continuity-based framework*, views the RD design as a misspecified regression model, local to the cutoff, and employs nonparametric smoothing techniques (Hahn et al., 2001). The differences and similarities between the local randomization and continuity-based frameworks for RD analysis are discussed in detail in Cattaneo et al. (2017), Cattaneo et al. (2020a, 2022), Cattaneo et al. (2020c), Cattaneo and Titiunik (2022), and Sekhon and Titiunik (2016, 2017).

In this article, we provide a systematic overview of recent statistical methodologies to analyze

and interpret RD designs employing both the continuity and the local randomization frameworks, with a focus on biomedical applications. Our discussion covers key assumptions, diagnostic tests, estimation methods, and inference procedures. While there are several resources available on the analysis of RD designs for the biomedical sciences (Bor et al., 2014; O'Keeffe et al., 2014; Maciejewski and Basu, 2020), these introductory articles do not discuss the most recent estimation and inference methods, which are already widely adopted in the statistical, social and behavioral sciences. Furthermore, available biomedical reviews do not address two complications that frequently arise in biomedical research: imperfect treatment compliance and non-continuous score variable.

The first complication, imperfect compliance, arises because in clinical applications RD designs are often based on recommended guidelines rather than enforceable rules. Many RD designs in medicine arise from clinical guidelines based on diagnostic scores. In these cases, the physician typically uses the available diagnostic information but do not follow the guidelines strictly. Instead, the physician, based on expert knowledge and additional patient-specific information, may decide to leave some patients above the cutoff untreated or treat some patients below the cutoff—or both. As we note below, one common pattern is that physicians tend to offer treatment to patients just below the clinical guideline. This results in imperfect compliance with the RD assignment rule, turning the design into a fuzzy RD setup where the probability of being treated jumps abruptly at the cutoff, but not necessarily from 0 to 1.

The fuzzy RD design is conceptually similar to an instrumental variables (IV) design in a local neighborhood around the cutoff c, where the score acts as an instrument for exposure to the treatment for units whose score is sufficiently close to c. This equivalence has two important conceptual implications. First, RD designs based on clinical guidelines can be thought of as a class of IV designs, at least locally to the cutoff, much in the same way as sharp RD designs can be thought of as local randomized experiments. From this perspective, fuzzy RD designs add a new class of instruments to medical research, enlarging the set of commonly used IVs in biomedical research (e.g., physician treatment preferences, distance to medical facilities, and genetic variants). Second, the analysis of fuzzy RD designs based on clinical applications requires the careful combination of both RD and IV methods. Crucially, fuzzy RD designs lead to treatment effects that (though related) are conceptually different from those in standard IV designs. Estimation and inference

also needs to be adjusted to account for specific RD features. As a result, the best practices for analysis and interpretation require both RD and IV methods. Below we outline how these methods should be used together, which is a novel feature of this tutorial relative to the extant literature.

The second complication encountered in medical applications is related to score coarseness. In the classical RD design, the score is assumed to be a continuous variable. Theoretically, the score is continuous when the set of values that it can take contains an uncountable number of elements. In practice, the score is considered continuous when the observations in the dataset have distinct values—that is, when there are no ties or repeats in the score values. In settings where the RD score is, for example, a poverty index or test score, this assumption is usually satisfied. However, in many applications, the RD score is discrete and thus exhibits values that are shared by more than one observation or *mass points*. Even the mechanical rounding of a continuous score can lead to RD designs with mass points (Dong, 2015), a problem sometimes known as *heaping* (Barreca et al., 2016). Sometimes the score takes on a relatively large number of mass points, while other times it exhibits only a few unique values.

The correct analysis and interpretation of the RD design depends crucially on whether the score is continuous or not, and on how many distinct values the score takes when it is not continuous. Most methods for the analysis of RD designs assume the score is continuous, and indeed prior RD tutorials in medicine have exclusively focused on that specific setting. However, non-continuous scores are common in biomedical research, as we exemplify below with a re-analysis and discussion of three medical applications. As our applications show, discrete scores and their coarseness can arise for many reasons and take different forms. RD designs with non-continuous scores require special care when analyzed using modern methods, some of which were developed for continuous scores only; in some cases the same methods can be used but their interpretation changes, while in other cases specific methods for RD methods designed with discrete scores may be more appropriate. We discuss these issues in detail and illustrate them with empirical examples.

The rest of the manuscript is organized as follows. In the next section, we present three clinical examples that we use throughout the manuscript to illustrate the main methods and ideas discussed. Each application exemplifies a key aspect of RD designs that affects both the research design and the subsequent analysis. After introducing the applications, in Section 3 we present the general RD setup and introduce the notation. We then proceed to discuss specific methods for

the analysis and interpretation of the RD design, a discussion that we divide in two parts. First, in Section 4, we discuss how to analyze RD designs when the score is approximately continuous and most observations have a unique score value. Then, in Section 5, we discuss how to analyze RD designs when the score is instead discrete and there are many observations that share the same score value. Each one of these sections discusses estimation and inference in the continuity-based framework, estimation and inference in the local randomization framework, methods to evaluate validate the RD assumptions, and an empirical illustration of the methods. Section 6 concludes. General-purpose software and replication files in `R`, `Stata`, and `Python` are available at https://rdpackages.github.io/.

2 Three RD Design Applications in the Biomedical Sciences

We review the three clinical applications we use to motivate concepts and demonstrate methods. In the first application, the RD score is discrete but can be treated as approximately continuous because there are many distinct score values and in fact most values of the score have only one or two observations. In the second application, the score is discrete with a few distinct values and many observations for each value of the score. In this application, it is infeasible to treat the score as approximately continuous. Both applications are fuzzy RD designs. The third application has discrete scores with many distinct values and thus can be analyzed using both continuity-based and local randomization methods.

2.1 Application 1: CD4 Counts and Anti-retroviral Therapy

Our first application is based on the Hlabisa HIV Treatment and Care Programme in South Africa, a study conducted by the Africa Health Research Institute (https://www.ahri.org/) and the South African Department of Health. The 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 (Tanser et al., 2007; Houlihan et al., 2010). Patients were eligible for anti-retroviral therapy (ART) if their CD4 count was less than 350 cells/μl, and they had a WHO stage III/IV condition (Bor et al., 2017). 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.

We re-analyze a recent study that used an RD design to estimate the effect of immediate (versus deferred) ART on retention in care (Bor et al., 2017). In that analysis, the investigators compared differences in retention between patients presenting with CD4 counts just above versus just below the 350-cells/μl threshold. The cohort included 11,306 patients and the data includes information on several predetermined covariates, including sex, age, date of testing, and testing location. This is an RD design where the unit of observation is the patient, the running variable is the patient's CD4 count, the cutoff is 350, the treatment is the immediate initiation of ART, and the outcome of interest is an indicator for 12-month retention in care. This indicator is 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. Henceforth, we refer to this empirical application as the *ART* application.

2.2 Application 2: Genetic Guidelines for Chemotherapy

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 (Paik et al., 2006; Sparano and Paik, 2008; Albain et al., 2010).

For this application, we analyze a cohort of patients from the Penn Breast Database from 2009 to 2017. We analyze patients 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 3,269. We also exclude 3 patients who did not undergo oncotype

scoring and those who did not undergo surgery for a final cohort of $3,224$ patients. 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, the running variable is the patient's oncotype score, the cutoff is 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 also fuzzy since adjuvant chemotherapy was prescribed to patients with scores of less than 26. Henceforth, we refer to this empirical application as the *chemotherapy* application.

2.3 Application 3: Patient Cost-Sharing and Healthcare Utilization

In many countries, health care costs are subsidized through government programs. 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. For example, health care for children is free or subsidized at higher rates than for adults. Variation in cost-sharing by age creates a discontinuity in health care subsidization, which can be exploited as an RD design. 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 re-analyze the study by Han et al. (2020), who studied this question in Taiwan, where, as of 2020, all inpatient and outpatient services for children under the age of 3 are completely exempt from copayments. They used this discontinuity in age to compare levels of health care utilization just before and after the third birthday.

The data includes 414,282 children born between 2003 and 2004. In its original form, the score is discrete, defined as the number of days until the child's third birthday—normalized to be zero on the day of the third birthday. In the original analysis, the score was further coarsened, since individual-level data was further collapsed into age cells measured in days. 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 greater than 3.

This captures the higher level of patient's cost-sharing due to the expiration of the subsidy after the third birthday. Unlike our previous two applications, this RD design is sharp, since once the child is 3 years of age or older there are no exceptions to the change in cost-sharing. The original study examined several outcome measures of healthcare utilization. We focus on one of those measures of utilization: the number of medical visits per 10,000 person days. This is an RD design where the unit of observation is the child but the data are collapsed to the day-level, the outcome is the number of medical visits per 10,000 person days, the score is the number of days between the day when the outcome is measured and the child's third birthday, the cutoff is normalized to zero, and the treatment is the elimination of healthcare subsidies. Henceforth, we refer to this study as the *cost-sharing* application.

3 RD Setup

A RD design is a study where each unit receives a *score*—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: 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. 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 feature contributes to the RD design's superior credibility when compared to other non-experimental methods for the analysis of observational data.

We assume that there are n units, indexed by $i = 1, 2, \ldots, n$, and each unit has a score value X_i. There is a single, common known cutoff c, and all units with $X_i \geq c$ are assigned to the active treatment condition, while all units with $X_i < c$ are assigned to the control condition.[1] The assignment rule $T_i = \mathbb{1}(X_i \geq c)$ implies that 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.

[1] This canonical design can be generalized to the case of multiple scores (Papay et al., 2011; Reardon and Robinson, 2012), geographic RD design (Keele and Titiunik, 2015; Keele et al., 2015), multiple cutoffs (Cattaneo et al., 2016, 2021), among other possibilities. See Cattaneo and Titiunik (2022) for more examples and references.

However, being *assigned* to the treatment condition is not the same as actually *receiving* the treatment. This is a critical distinction that defines the difference between sharp and fuzzy RD designs. In the sharp RD design, treatment assignment is identical to treatment received for all units, while in the fuzzy RD design the two differ for some units. We pay special attention to fuzzy RD designs because, as we noted above, compliance is likely to be imperfect in many RD designs in clinical settings where physicians may sometimes act contrary to recommended guidelines based on diagnostic tests. We use the (binary) variable D_i to denote whether the treatment was actually received by unit i. Using this notation, we see that in the sharp RD design we always have $T_i = D_i$ because compliance with treatment assignment is perfect, while the defining feature of the 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 ($T_i = \mathbb{1}(X_i \leq 350) = 1$) who never initiate ART ($D_i = 0$).

We adopt the potential outcomes framework (Imbens and Rubin, 2015; Hernán and Robins, 2022) and assume that each unit has one outcome corresponding to each possible value of the treatment assignment and the treatment received, and one treatment received corresponding to each value of the treatment assigned. For the treatment received, this means that 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 (i.e, when $X_i \geq c$ and $T_i = 1$), while $D_i(0)$ is the treatment that unit i receives when this unit is assigned to the control condition (i.e, when $X_i < c$ and $T_i = 0$). Both $D_i(1)$ and $D_i(0)$ can be one or zero, depending on whether 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 $D_i(0) = 0$.

For the outcome of interest, this framework implies that every unit has four different potential outcomes depending on the combination of the treatment assignment and the compliance decisions. 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 had score below the cutoff ($T_i = 0$) but received the treatment anyway ($D_i(0) = 1$). We have four potential outcomes in total: $Y_i(1, 0)$, $Y_i(1, 1)$, $Y_i(0, 0)$ and $Y_i(0, 1)$. 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, we write the observed treatment received as

$$D_i = (1 - T_i) \cdot D_i(0) + T_i \cdot D_i(1) = \begin{cases} D_i(0) & \text{if } T_i = 0 \\ D_i(1) & \text{if } T_i = 1 \end{cases},$$

and the observed outcome as

$$Y_i = (1 - T_i) \cdot Y_i(0, D_i(0)) + T_i \cdot Y_i(1, D_i(1)) = \begin{cases} Y_i(0, D_i(0)) & \text{if } T_i = 0 \\ Y_i(1, D_i(1)) & \text{if } T_i = 1 \end{cases}.$$

Finally, we assume the observed data is $(Y_1, D_1, X_1), \ldots, (Y_n, D_n, X_n)$. In most of our discussion, we assume that the observations are a random sample from a larger population and the potential outcomes are therefore random variables. We deviate from this setup only when discussing Fisherian inference, which assumes that the potential outcomes are non-stochastic and hence the observed outcomes are random only because of the randomness induced by the treatment assignment mechanism.

We use the notation above to define various treatment effects in terms of comparisons between features of (the distribution of) potential outcomes. Ideally, we would like to study treatment effects for all units, 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 effect is not generally available in RD designs because the non-experimental treatment assignment will only justify studying effects for units whose scores are near the cutoff. Furthermore, non-compliance and/or discrete score will further limit the type of RD treatment effects that can be learned from data.

4 RD Analysis when the Score is Continuous

We start by discussing interpretation, estimation, and inference in RD designs where the score is either continuous (all or most of the observations have a unique value of X_i) or has a large number of mass points (multiple observations share the same value of the score, but there are many distinct such shared values). We first discuss how to illustrate the design graphically using RD plots, and then proceed to discuss continuity-based methods, followed by local randomization methods, and

then methods for validation of the RD assumptions. We conclude this section with an empirical illustration of the methods based on the ART application, which we also use as a running example throughout the section. In Section 5, we re-evaluate the appropriateness and effectiveness of the methods discussed in this section to settings where the score is discrete and exhibits few distinct values, and also introduce other approaches more suitable for that setting.

4.1 Illustrating the Design with RD Plots

A useful first step in RD analysis is a graphical illustration of the design (Calonico et al., 2015). When properly executed, a graphical RD analysis adds transparency and credibility by displaying the observations used for estimation and inference, both globally on the full support of the score and locally near the cutoff. RD plots are also useful to 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. Despite their usefulness, RD plots should only be employed for graphical presentation of the RD design and not for estimation and inference of RD treatment effects (Korting et al., 2022).

Although we could easily construct a raw scatter plot of the observed outcome against the score, such plot would likely hide many interesting features in the outcome-score relationship like discontinuities or non-linearities. For this reason, it is customary to "smooth" the data before plotting, which is usually done by binning or partitioning the support of the score and then reporting the average outcome within each bin.

When the score is continuous, RD plots are constructed by first choosing disjoint (i.e., non-overlapping) intervals or "bins" of the score, calculating the mean of the outcome for the observations falling within each bin, and then plotting the average outcome in each bin against the mid point of the bin. These binned means can be interpreted as a non-smooth approximation to the unknown regression functions. 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 fourth- or fifth-order polynomial of X_i using the original (i.e., not binned) data. The global polynomial fits can be interpreted as a smooth approximation of the unknown regression functions, in contrast to the non-smooth approximation provided by the bins.

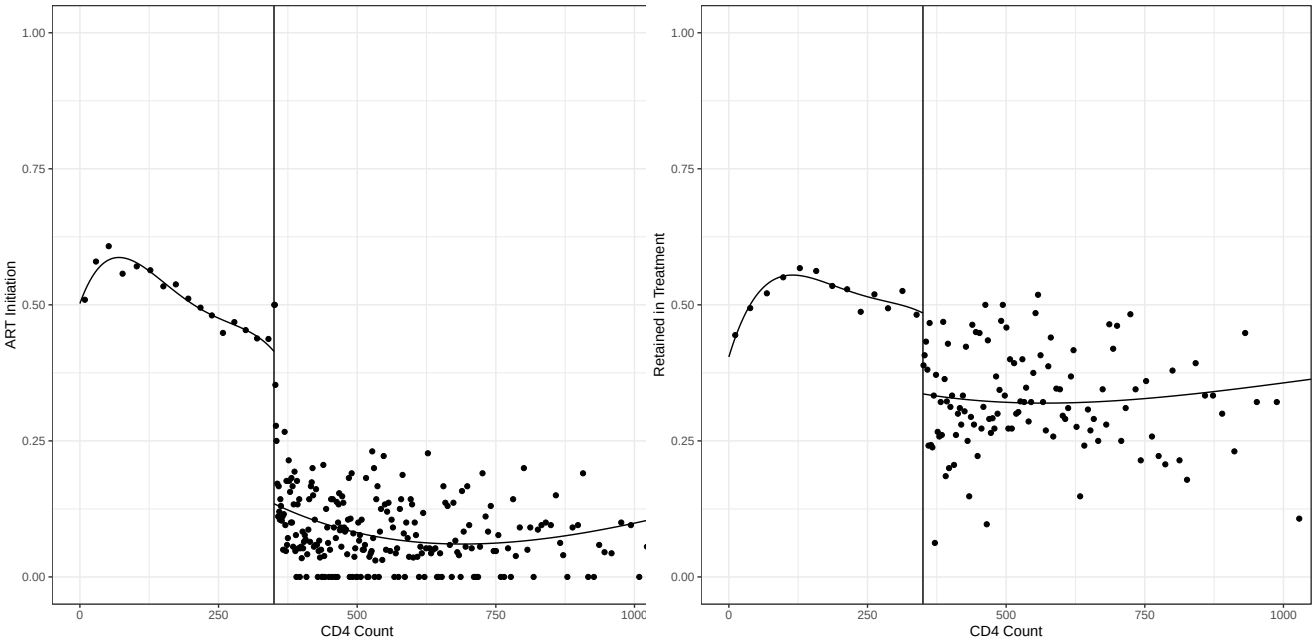
Choosing the appropriate global polynomial order is important. The global fit represents a

smooth approximation of the unknown regression functions: 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 well-known to be particularly detrimental at boundary points, which is the area of interest in RD designs. See Calonico et al. (2015), Gelman and Imbens (2019), and Korting et al. (2022) for further discussion, and Pei et al. (2021) for a recent discussion on polynomial order selection in RD designs.

An RD plot with smoothed bins and global polynomial fits provides a useful visualization of the overall shape of the regression functions for treated and control observations, while at the same time retains enough information about the local behavior of the data to observe the RD treatment effect and the variability of the data around the global fit. RD plots can be implemented in a variety of ways, depending on how the bins are chosen. Calonico et al. (2015) proposed several automatic data-driven bin selection methods and studied their statistical properties. The positions of the bins over the support of the score can be selected so that all bins have the same length (evenly-spaced placement) or all bins have the same number of observations (quantile-spaced placement), while the total number of bins can be selected to either maximize the approximation of the unknown regression function or to better represent the overall variability of the data. A commonly used RD plot uses evenly-spaced bin placement with a total number of bins selected to match the overall variability of the data.

For fuzzy RD designs, it is standard to provide two plots, one for the relationship between the score (X_i) and the observed outcome (Y_i), and the other for the relationship between the score (X_i) and the treatment taken (D_i). Figure 1 shows RD plots for the ART application, where the score X is a patients' CD4 count, the treatment assignment is $T_i = \mathbb{1}(X_i \leq 350)$, and the treatment received D_i is equal to one if the patient initiated ART. Figure 1a plots the average of D_i (the proportion of patients that started ART) within bins of CD4 count (X_i). The bins are quantile-spaced and chosen to approximate the unknown regression function; the global fit is of order four. The effect of being below the CD4 count threshold of 350 on ART initiation is clearly visible. Patients are more than 25% points more likely to start ART when their CD4 count is 350 or lower. In Figure 1b, we observe that patients below the CD4 count threshold of 350 are about 15% points more likely to be retained in treatment.

Figure 1: RD plots for ART application using quantile-spaced bins.



(a) Proportion with ART initiated against CD4 count.

(b) Proportion retained in treatment against CD4 count.

4.2 Continuity-Based Methods

We now discuss how to formally estimate and make inferences for the kind of effects illustrated graphically above. Our discussion begins with the sharp RD design, where treatment assignment and treatment received perfectly coincide. When compliance is perfect, the notation simplifies considerably because the treatment assigned and the treatment received coincide for all units ($D_i = T_i$), and the four potential outcomes reduce to two, $Y_i(1) \equiv Y_i(1,1)$ and $Y_i(0) \equiv Y_i(0,0)$. The most common causal estimand in the sharp RD design is the average treatment effect at the cutoff, formally defined as

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

where the expectation should be interpreted based on the causal inference framework used.

In the case of random potential outcomes, the expectation is taken with respect to the underlying (population) probability distribution. On the other hand, when the potential outcomes are non-random, the expectation becomes a sample average across all units in a finite-sample population or an integral with respect to some common feature underlying the definition of potential outcomes in

an infinite population (i.e., some form of ergodicity). In causal inference settings, these distinctions are usually associated with different statistical models, usually called *super-population*, *Neyman*, *Fisher*, or *finite-sample* (causal) models, among other possibilities. See Rosenbaum (2010), Imbens and Rubin (2015) and Hernán and Robins (2022) for more details. In the remainder of this article, we adopt a super-population model where the potential outcomes are regarded as random variables with a common probability law in the population. We only depart from this framework to discuss Fisherian methods for inference in Section 4.3.1.

The parameter $\tau_{\texttt{SRD}}$ is the sharp RD treatment effect for units with score equal to the cutoff. In the cost-sharing application, this parameter answers a specific counterfactual question: what would be the average number of medical visits for children who are three years old if we switched their status from health care subsidies (below the cutoff) to no health care subsidies (above the cutoff). It is critical to note that the effect recovered by the RD design is the average effect of treatment for units *local to the cutoff*—that is, for units with score values $X_i = c$. This implies that the RD treatment effect tends to have limited external validity, because it is generally not representative of the treatment effect that would occur for units with scores away from the cutoff (see Cattaneo et al., 2021, for more discussion and related references). In our example, the RD effect is for children who are exactly (or near) three years of age, but as we consider children who are considerably younger or older, we cannot extrapolate this effect without additional assumptions.

The valid estimation of $\tau_{\texttt{SRD}}$ is based on the key idea that units with similar values of the score but on opposite sides of the cutoff should be "comparable" in all aspects except the fact that units whose scores are above the cutoff were assigned to treatment while units whose scores are below the cutoff were not. In our example, we assume that children who just turned three years of age are similar in all important characteristics to those children who are a few days away from their third birthday. This assumption of comparability is based, more formally, on a continuity assumption that we now outline.

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. 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:

$$\tau_{\texttt{SRD}} = \lim_{x \downarrow c} \mathbb{E}[Y_i | X_i = x] - \lim_{x \uparrow c} \mathbb{E}[Y_i | X_i = x]. \tag{1}$$

In Equation (1), 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. That is, under continuity, children just above and below three years of age should have similar potential outcomes on average, which justifies using the difference in outcomes for these two groups as an estimate of the treatment effect.

We can extend these ideas to the fuzzy RD design, with some modifications. The most important difference is that the parameter $\tau_{\texttt{SRD}}$, which captures the average effect of the treatment received on all units with scores near the cutoff, is unavailable except under strong assumptions that will be implausible in many applications (e.g., constant treatment effects as a function of the score). Instead, when there is non-compliance, researchers typically focus on two types of 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, and is the effect of interest in many cases. For example, in the ART application we are interested in the effect of initiating ART on patient retention, and in the chemotherapy application we are interested in the effect of administering adjuvant chemotherapy on breast cancer recurrence. However, in some cases, researchers are also interested in the effect of assigning the treatment on the outcome, which are commonly known as *intention-to-treat* (ITT) effects. These effects include not only the effect that the treatment received may directly have on the outcome, but also effects caused by strategic decisions that individuals make in response to knowledge about their assignment. Policy-makers interested in anticipating the overall effects of establishing a new program will typically be interested

in ITT effects.

We start by considering the effects assigning the treatment on both the outcome (Y_i) and the treatment received (D_i). A sharp RD analysis of the effect of treatment assignment T_i on the outcome Y_i estimates the parameter defined as

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

Under appropriate continuity assumptions analogous to those in the sharp RD case, $\tau_{\texttt{Y}}$ captures the ITT effect of the treatment assignment on the outcome at the cutoff, which we can write as $\tau_{\texttt{Y}} = \mathbb{E}[Y_i(1, D_i(1)) - Y_i(0, D_i(0)) | X_i = c]$, the average change in potential outcomes at the cutoff from switching the assignment from control to treated.

In other words, 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, when interest is on the effect of the treatment assignment, 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 the sharp RD settings, but we modify the assumptions and interpretation to accommodate the presence of imperfect compliance. Because some units fail to comply with their assignment, the effect of T_i on Y_i at the cutoff is no longer the effect of the treatment itself, but rather the effect of *assigning* the treatment to the units in the analysis. Unfortunately, the average effect of the treatment at the cutoff, $\tau_{\texttt{SRD}}$, is not generally available in the fuzzy RD design. For example, in the ART application, $\tau_{\texttt{Y}}$ captures the average effect of offering ART to patients whose CD4 is 350 who may or may not accept the offer, while the parameter $\tau_{\texttt{SRD}}$ captures the effect of actually starting ART for those patients.

When continuity conditions hold, the sharp RD effect, $\tau_{\texttt{SRD}}$, and the ITT RD effect, $\tau_{\texttt{Y}}$, are equivalent in the particular case when all units are compliers, that is, when $D_i(1) - D_i(0) = 1 - 0 = 1$ for all i. In this case, the effect of assigning the the treatment at the cutoff is equivalent to the effect of receiving the treatment at the cutoff. In contrast, in the general case when $D_i(1) - D_i(0) \neq 1$, the two parameters differ.

In the fuzzy RD setting, we can further investigate the effect of the treatment assignment, T_i, on the treatment received, D_i, at the cutoff. The sharp RD estimator of the effect T_i on D_i estimates

the parameter $\tau_{\mathtt{D}}$, defined as

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

Since D_i is binary, $\tau_{\mathtt{D}}$ captures the difference in the probability of receiving the treatment at the cutoff between units assigned to the treatment vs. the control condition. In our application, this is the difference between the proportion of patients with CD4 counts above 350 who initiate ART and the proportion of patients with CD4 counts below 350 who initiate ART.

Under continuity conditions, the difference in treatment probabilities captured by $\tau_{\mathtt{D}}$ can be attributed to the RD assignment rule; in this case, $\tau_{\mathtt{D}}$ represents the average effect of assigning the treatment on receiving the treatment at the cutoff, that is, $\tau_{\mathtt{D}} = \mathbb{E}[D_i(1) - D_i(0) | X_i = c]$. This effect is usually called the *first-stage* or *take-up* effect. Both $\tau_{\mathtt{Y}}$ and $\tau_{\mathtt{D}}$ are sharp RD parameters.

Although the ITT parameters $\tau_{\mathtt{Y}}$ and $\tau_{\mathtt{D}}$ are of independent interest in many applications, investigators are often primarily interested in the effect of receiving the treatment, not merely of assigning it. As mentioned above, $\tau_{\mathtt{SRD}}$ is infeasible in the fuzzy RD design due to non-compliance. But under some additional assumptions, it is possible to estimate a related parameter that captures the average effect of the treatment at the cutoff, at least for a particular subpopulation. We define the following parameter,

$$\tau_{\mathtt{FRD}} \equiv \frac{\tau_{\mathtt{Y}}}{\tau_{\mathtt{D}}},$$

which we call the fuzzy RD effect, and 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 $\tau_{\mathtt{FRD}}$ is of interest because it 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; a thorough discussion is beyond the scope of our discussion, but we refer the reader to Dong (2018), Cattaneo et al. (2016), and Arai et al. (2021a) for examples, and to Cattaneo et al. (2022) for a practical discussion.

Regardless of the technical details, interpreting the fuzzy RD parameter as the treatment effect

for compliers requires several assumptions. The formalization of these assumptions varies depending on the particular definitions adopted, but the conceptual ideas are similar in most cases. First, the parameter $\tau_{\mathtt{D}}$ must be nonzero (and well-separated from zero) for estimation and inference to be meaningful. In other words, being above (or below) the cutoff must induce some units to actually take the treatment. This rules out, for example, that having a CD4 count below 350 induces no patient to start ART. This is usually referred to as the *relevance* condition or the *first-stage* in IV settings.

Second, we need continuity conditions similar to those invoked in the sharp RD case, but generalized to cover the more complex setting of non-compliance. 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, crossing the cutoff should only affect the outcome if it has the 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. A medical treatment has a placebo effect when the patient experiences an improvement in symptoms that is attributable to the cues experienced during her participation in the therapeutic encounter (labels, interactions with clinicians, etc.) and not to the actual therapy received (Kaptchuk and Miller, 2015). When placebo effects are expected, the continuity conditions required in the fuzzy RD setting will not hold, because the treatment assignment will have an independent effect on the outcome, which will induce a discontinuity of the underlying regression functions at the cutoff. Thus, in applications where placebo effects are expected, researchers should be careful about the interpretation of $\tau_{\mathtt{FRD}}$ and possibly restrict their analysis to consider only ITT effects (which will still be valid, if perhaps difficult to interpret, in the presence of placebo effects).

In our ART example, the exclusion restriction requires that the CD4 count 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 communicate with patients differently, which can in turn lead to discovery of other health issues and return for care other than ART (recall that the outcome of interest in this application is retention in care). The exclusion restriction would be more plausible if the outcome were a biological manifestation of

HIV, 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 a *monotonicity* condition (Imbens and Rubin, 2015). 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 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 $x > c$, and a unit who takes the treatment when the cutoff is c must also take the treatment for any cutoff $x < c$. 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 360.

Although the particular formalization of these assumptions may vary, the important point is that, in the fuzzy RD design, the study of treatment effects requires more assumptions than in the sharp RD case. The necessity of additional assumptions underscores a key point: if the effect of D on Y is of interest, then a fuzzy RD design is a marriage of an RD design and an IV design. As such, a fuzzy RD design requires key ideas, assumptions and methods from both designs. In fact, given the pervasive nature of the fuzzy RD design in biomedical applications, one might fruitfully label these RD designs as a class of instruments.

4.2.1 Continuity-based Estimation and Inference

In recent years, much progress has been made in the development of continuity-based methods for estimation and inference in RD designs. This section focuses on the most popular approach in the literature: local polynomial estimation methods (Fan and Gijbels, 1996; Hahn et al., 2001) coupled with point-estimation-optimal bandwidth selection and robust bias correction inference (Calonico et al., 2014, 2019, 2020). See also Cattaneo and Vazquez-Bare (2016) and Calonico et al. (2020) for an overview on neighborhood selection methods in RD designs, and Calonico et al. (2018, 2022), Kamat (2018) and Tuvaandorj (2020) for theoretical results on robust bias correction methods.

When the score variable is continuous, a key problem for most estimation and inference methods in RD designs is that there are often a few observations with score arbitrary close to the cutoff. As such, estimates must rely on some form of local extrapolation, because estimating the effect at $X_i = c$ requires using observations whose values of X_i are not equal to c—i.e., observations

"away from the cutoff." In this case, estimation and inference proceeds by first *approximating* the unknown regression functions of potential outcomes, and then computing the estimated treatment effect and/or the statistical inference procedure of interest. Because a sufficiently smooth function can be well approximated by a polynomial function, locally or globally, up to an error term, the unknown regression functions, $\mathbb{E}[Y_i(0)|X_i = x]$ and $\mathbb{E}[Y_i(1)|X_i = x]$, can be approximated by a polynomial function of the score, at least locally to the cutoff.

Early empirical work in RD designs employed global methods with flexible higher-order polynomials, usually of 4th or 5th order, over the entire support of the data. However, the RD point estimator is defined at a boundary point, and global polynomial methods can lead to unreliable RD point estimates (i.e., Runge's phenomenon). Modern RD estimation uses local polynomial methods that discard observations sufficiently far away from the cutoff and employ only a low-order polynomial approximation (usually linear or quadratic) for estimation. This approach is more robust and less sensitive to boundary and overfitting problems. The standard, default approach is to employ two linear polynomial fits using only using observations near the cutoff point, separately for control and treatment units.

In particular, this approach uses only observations that are between $c - h$ and $c + h$, where $h > 0$ is the bandwidth that determines the size of the neighborhood around the cutoff. Within this neighborhood, it is common to adopt a weighing scheme to ensure that the observations closer to c receive more weight than those further away; the weights are determined by a kernel function $K(\cdot)$. The two preferred options for kernel weighting scheme are the triangular kernel, $K(x) = (1 - |x|)\mathbb{1}(|x| \leq 1)$, which linearly down-weights observations within the neighborhoods $[c - h, c)$ and $[c, c + h]$, and the uniform kernel, $K(x) = \mathbb{1}(|x| \leq 1)$, which puts equal weight to observations within that neighborhood. From a technical perspective, the local polynomial approach is understood as a nonparametric regression fit that approximates the unknown underlying regression functions within the region determined by the bandwidth.

To summarize, standard local polynomial estimation for RD designs consists of the following basic steps:

1. Choose a polynomial order p and a kernel function $K(\cdot)$. Defaults are linear fit ($p = 1$) and triangular kernel ($K(x) = (1 - |x|)\mathbb{1}(|x| \leq 1)$).

2. Choose a bandwidth h. Default is a mean squared error optimal choice (see below for more details).

3. For observations above the cutoff (i.e., 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 $K(\frac{X_i - c}{h})$ for each observation. The estimated intercept from this local weighted regression, $\widehat{\mu}_+$, is an estimate of the point $\mu_+ = \mathbb{E}[Y_i(1)|X_i = c]$:

$$\widehat{\mu}_+ \quad : \quad \widehat{Y}_i = \widehat{\mu}_+ + \widehat{\mu}_{+,1}(X_i - c) + \widehat{\mu}_{+,2}(X_i - c)^2 + \cdots + \widehat{\mu}_{+,p}(X_i - c)^p.$$

4. For observations below the cutoff (i.e., observations with $X_i < c$), 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 $K(\frac{X_i - c}{h})$ for each observation. The estimated intercept from this local weighted regression, $\widehat{\mu}_-$, is an estimate of the point $\mu_- = \mathbb{E}[Y_i(0)|X_i = c]$:

$$\widehat{\mu}_- \quad : \quad \widehat{Y}_i = \widehat{\mu}_- + \widehat{\mu}_{-,1}(X_i - c) + \widehat{\mu}_{-,2}(X_i - c)^2 + \cdots + \widehat{\mu}_{-,p}(X_i - c)^p.$$

5. Calculate the sharp RD point estimate:

$$\widehat{\tau}_{\texttt{SRD}} = \widehat{\mu}_+ - \widehat{\mu}_-.$$

Local polynomial methods require the user to make three choices: the polynomial order, the kernel function, and the bandwidth. As mentioned above, it is standard to employ a linear polynomial and a triangular kernel, as these choices have objective theoretical advantages in the nonparametric literature. Moreover, it is common to investigate the robustness of the empirical results by choosing $p = 2$ or a uniform kernel. It is rare and not advisable, however, to employ higher-order polynomials: $p \geq 3$ is regarded as the largest value recommended in most RD settings. In practice, estimation results are typically insensitive to the choice of kernel, and to a lesser extent to the choice of p. In most settings, $p = 1$ or $p = 2$ are the only reasonable choices, with $p = 1$ the most natural default.

The bandwidth controls the width of the neighborhood around the cutoff that is used to fit the local polynomial models. That is, the bandwidth determines the number of observations above and

below the cutoff that are used for estimation. As is true with any nonparametric regression model, a small bandwidth will reduce the approximation error (also known as *smoothing bias*) of the local polynomial approximation, since it only uses observations very close to the cutoff. However, a small bandwidth will increase the variance of the estimated coefficients 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.

Bandwidth selection can be automated in a principled, data-driven way following the bias-variance trade-off. The most natural approach is to select a bandwidth that minimizes an approximation to the mean squared error (MSE) of the RD point estimator, $\widehat{\tau}_{\texttt{SRD}}$. The MSE of any estimator is the sum of its bias squared plus its variance. As such, if one uses the MSE to select h, one is selecting h to optimize the bias-variance trade-off. Informally, these methods derive the asymptotic MSE approximation for the RD estimator, and then choose the value of h that minimizes it. This MSE-optimal bandwidth selection approach has become the standard for RD estimates. However, it is important to note that the MSE-optimal bandwidth is optimal for point estimation, but *invalid* for inference in general (Calonico et al., 2014, 2019, 2020).

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, by construction, even in large samples, which leads to confidence intervals and hypothesis tests that are invalid in general. This implies that the usual asymptotic 95-percent confidence interval for $\tau_{\texttt{SRD}}$ given by $\texttt{I} = \left[\widehat{\tau}_{\texttt{SRD}} \pm 1.96 \cdot \sqrt{\widehat{\texttt{V}}}\right]$, where $\widehat{\texttt{V}}$ denotes a variance estimator, is invalid because the underlying Gaussian distribution of the RD point estimator has a non-zero bias when the MSE-optimal bandwidth is used. In short, standard confidence intervals (CI) depend on the unknown misspecification bias, denoted $\texttt{B}$, which leads to undercover zero (or, by duality, to over-reject the null hypothesis of no-treatment effect, $\tau_{\texttt{SRD}} = 0$).

The bias term $\texttt{B}$ arises because of the very nature of the RD design: the local approximation is misspecified in general, which leads to a bandwidth selection that tries to balance bias and variance near the cutoff. Of course, if the local parametrization were correct, that is, if $\mathbb{E}[Y_i(0)|X_i = x]$ and $\mathbb{E}[Y_i(1)|X_i = x]$ were exactly linear functions near the cutoff and a local linear polynomial is used, then $\texttt{B} = 0$ and inference based on $\texttt{I}$ would be valid. However, in such case, bandwidth

selection would be invalid. More importantly, in arguably all RD applications based on continuity assumptions, the functional forms of $\mathbb{E}[Y_i(0)|X_i = x]$ and $\mathbb{E}[Y_i(1)|X_i = x]$ are unknown and hence must be approximated, leading to a non-zero misspecification bias ($\mathtt{B} \neq 0$) and the need for bandwidth selection in the first place. This is why the term $\mathtt{B}$ appears in the distributional approximation when nonparametric methods are employed, but not when the method of estimation is regarded as parametric.

Investigators can choose to ignore the bias, potentially leading to false discoveries of non-zero treatment effects, or choose a bandwidth smaller than the MSE-optimal choice when forming the conventional CI—a procedure known as "undersmooothing." However, undersmoothing is not recommended because there are no clear criteria for shrinking the bandwidth, which can result in *ad-hoc* decisions, lack of transparency, and specification searching. Moreover, undersmoothing reduces statistical power because a smaller bandwidth uses fewer observations for inference by construction.

A principled alternative is to use the *robust bias corrected* confidence intervals originally proposed by Calonico et al. (2014), and later extended to other settings by Xu (2017), Arai and Ichimura (2018), Calonico et al. (2019), Dong et al. (2021), and Arai et al. (2021b), among many others. See Calonico et al. (2020) for an overview and more references. These confidence intervals are based on a bias correction that first estimates the bias term $\mathtt{B}$ with an estimator $\widehat{\mathtt{B}}$ based on a higher-order polynomial fit, and then removes this term from the RD point estimator. The derivation of these robust confidence intervals allows the estimated bias term to converge in distribution to a random variable and thus contribute to the distributional approximation of the RD point estimator. This results in an asymptotic variance of the form $\mathtt{V} + \mathtt{W}$, which is larger than the original variance $\mathtt{V}$ used by the conventional approach. This increase in variability $\mathtt{W}$ captures, heuristically, the additional uncertainty introduced by the bias correction needed to account for the misspecification error due to the local polynomial approximation used in the first place.

Consequently, a robust bias corrected confidence interval is one that not only estimates the leading misspecification error but also incorporates the contribution of the bias correction step to the variability of the resulting statistic. This approach leads to robust bias corrected confidence

intervals of the form

$$\mathtt{I}^{\mathtt{rbc}} = \left[\ \left(\widehat{\tau}_{\mathtt{SRD}} - \widehat{\mathtt{B}}\right) \pm 1.96 \cdot \sqrt{\widehat{\mathtt{V}} + \widehat{\mathtt{W}}} \ \right],$$

which are constructed by recentering the CI with the bias estimate ($\widehat{\mathtt{B}}$) and simultaneously rescaling it with a new variance formula, where $\widehat{\mathtt{W}}$ is an estimator of the correction term $\mathtt{W}$ that accounts for the added variability in the bias correction step. These confidence intervals are centered around the bias corrected point estimate, not around the original estimate $\widehat{\tau}_{\mathtt{SRD}}$.

Robust bias corrected 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 (Calonico et al., 2018, 2022; Kamat, 2018; Tuvaandorj, 2020). Furthermore, the good finite sample performance these intervals has been validated empirically in Ganong and Jäger (2018), Hyytinen et al. (2018) and De Magalhães et al. (2020), among others. As a consequence, for sharp RD analysis, we recommend (i) reporting the MSE-optimal RD point estimate $\widehat{\tau}_{\mathtt{SRD}}$, which is constructed using an MSE-optimal bandwidth choice, and (ii) reporting robust bias corrected confidence intervals $\mathtt{I}^{\mathtt{rbc}}$, which employ the same MSE-optimal bandwidth choice. All these methods are readily available in general-purpose software packages in `Stata`, `R`, and `Python`.

The local polynomial methods we have discussed so far for sharp RD continuity-based analysis can also be applied to fuzzy RD designs to estimate $\tau_{\mathtt{Y}}$, $\tau_{\mathtt{D}}$, and $\tau_{\mathtt{FRD}}$. The first point estimator can be constructed exactly as described above, using local polynomials to estimate the relationship between Y_i and the score X_i, since $\widehat{\tau}_{\mathtt{Y}} = \widehat{\tau}_{\mathtt{SRD}}$. The estimator $\widehat{\tau}_{\mathtt{D}}$ of $\tau_{\mathtt{D}}$ is constructed analogously, after replacing the observed outcome variable Y_i with the observed treatment status D_i. Once $\widehat{\tau}_{\mathtt{Y}}$ and $\widehat{\tau}_{\mathtt{D}}$ are available, the fuzzy RD estimand $\tau_{\mathtt{FRD}}$ is estimated using

$$\widehat{\tau}_{\mathtt{FRD}} = \frac{\widehat{\tau}_{\mathtt{Y}}}{\widehat{\tau}_{\mathtt{D}}},$$

the ratio of the estimated effect of T_i on Y_i and the estimated effect of T_i on D_i at the cutoff.

The estimator $\widehat{\tau}_{\mathtt{FRD}}$ is consistent for $\tau_{\mathtt{FRD}}$ 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 applies, 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 $\widehat{\tau}_{\texttt{FRD}}$, or two distinct bandwidth choices, one each for the numerator and denominator. In practice, most researchers employ a single MSE-optimal choice for $\widehat{\tau}_{\texttt{FRD}}$ or for $\widehat{\tau}_{\texttt{Y}} = \widehat{\tau}_{\texttt{SRD}}$, although some researchers prefer to choose two different bandwidths for $\widehat{\tau}_{\texttt{Y}}$ and $\widehat{\tau}_{\texttt{D}}$. As a general rule, it is usually recommended to use a single MSE-optimal bandwidth for the estimator of interest, in this case, $\widehat{\tau}_{\texttt{FRD}}$.

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. The resulting confidence interval takes exactly the same form as above, $\texttt{I}^{\texttt{rbc}}$, but now with $\widehat{\tau}_{\texttt{SRD}}$ replaced with $\widehat{\tau}_{\texttt{FRD}}$ and, by implication, with the associated changes in the bias and variance formulas. Because these formulas are cumbersome we do not reproduce them here, but they can all be found in Calonico et al. (2014), Calonico et al. (2019), and Calonico et al. (2020).

Lastly, we note that MSE-optimal bandwidth choices are not the only way of implementing local polynomial methods for the analysis of RD designs. Calonico et al. (2020) discuss other optimal bandwidth choices that are specifically tailored towards inference—in particular, confidence interval construction using robust bias correction methods. This alternative bandwidth selector is called CE-optimal because it is developed to minimize the coverage error (CE) of the robust bias corrected confidence intervals $\texttt{I}^{\texttt{rbc}}$. A CE-optimal bandwidth choice is appropriate for inference, including falsification testing or diagnostic of RD designs as discussed in Section 4.4. We do not discuss further this alternative bandwidth selection approach for brevity, but we note that those methods are also available in general-purpose software packages mentioned in the introduction. Section 4.5 illustrates how the continuity-based methods are deployed to the ART application.

4.3 Local Randomization Methods

In the continuity-based framework for RD designs, the analysis is conducted at the cutoff in a "limiting" sense. Consequently, treatment effects are defined as limiting quantities towards the cutoff point (see equation (1) and the definitions of $\tau_{\texttt{D}}$ and $\tau_{\texttt{Y}}$) and all estimation and inference

methods rely on large sample approximations where the neighborhood around the cutoff is assumed to shrink as the sample size increases. In other words, identification, estimation and inference of RD treatment effects crucially rely on some form of extrapolation to the cutoff, when the RD design analysis is based on continuity assumptions.

The local randomization framework for RD designs is conceptually different, offering a complementary set of methods for analysis and interpretation. The core idea is that there exists a neighborhood or window around the cutoff where the treatment assignment resembles what it would have been in an experiment. This idea, which was first discussed heuristically in the seminal RD paper of Thistlethwaite and Campbell (1960) and later popularized by Lee (2008), and was first formalized by Cattaneo et al. (2015) and later expanded by Cattaneo et al. (2017). Building on this work, the idea has been extended and applied to other RD contexts, including geographic RD designs (Keele et al., 2015), principal stratification (Li et al., 2015), and kink RD designs (Ganong and Jäger, 2018). See Sekhon and Titiunik (2016, 2017) for conceptual and methodological discussion about the advantages and limitations of the RD local randomization framework, and Cattaneo et al. (2022) for a practical textbook introduction.

Once the window around the cutoff is chosen, the operative assumption is that units with scores in this window have a treatment assignment mechanism resembling an experiment, which enables the deployment of standard methods from the analysis of experiments literature. The particular implementation of these tools depends on how the idea of local randomization is formalized, which can include finite-sample and large-sample perspectives. We cover two local randomization frameworks: (i) random potential outcomes and large samples, following the super-population model already used to discuss the continuity-based framework, and (ii) fixed potential outcomes and finite samples, following the Fisherian causal inference model. Due to space constraints, we omit discussion of the Neyman framework, which can be understood an intermediate case between the Fisherian and the super-population frameworks. See Cattaneo and Titiunik (2022) and Cattaneo et al. (2022) for further discussion in the context of RD designs, and Rosenbaum (2010), Imbens and Rubin (2015), and Hernán and Robins (2022) for textbook introduction to the three causal inference frameworks for the analysis of experiments.

Regardless of the framework used for analysis, the key feature is that the window around the cutoff where local randomization is assumed to hold is fixed even when conducting approximations

in large samples. If the neighborhood were shrinking with the sample size, then the analysis and interpretation of the methods would be as in the continuity framework introduced above, which is possible, but conceptually different from the standard local randomization interpretation. In practice, such distinction becomes less relevant because the local randomization RD framework will typically be valid only in "small" neighborhoods around the RD cutoff; we will return to this point below when discussing estimation and inference in that framework.

To formalize the local randomization RD framework, we retain the notation introduced in the previous section, but with the caveat that the potential outcome variables are assumed to be non-random whenever we discuss the Fisherian framework. In addition, we introduce notation for the local randomization neighborhood or window, $W = [c - w, c + w]$, where $w > 0$ is its half length and W is assumed symmetric around the cutoff only for simplicity. In this setting, we call W a *window* rather than a *bandwidth*, to distinguish it from the local neighborhood used in the context of continuity-based methods. Finally, we employ vector notation as follows. Bold variables without subscript correspond to the vector collecting all n variables, for example $\mathbf{Y}(\mathbf{0}) = (Y_1(0), Y_2(0), \ldots, Y_n(0))'$, $\mathbf{Y}(\mathbf{1}) = (Y_1(1), Y_2(1), \ldots, Y_n(1))'$, $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_n)'$, $\mathbf{T} = (T_1, T_2, \ldots, T_n)'$, etc. Bold variables with subscript W collect the subvector of the $n_W < n$ observations whose score satisfies $X_i \in W$. For example, the vector $\mathbf{Y}_W = (Y_i : X_i \in W)$ collects the observed outcomes only for those observations whose scores X_i lie within the window W, i.e. with $c - w \leq X_i \leq c + w$.

With the above notation, we can discuss the two key assumptions in the local randomization framework for RD designs: knowledge of the treatment assignment mechanism inside the window, and no direct effect of the running variable on the potential outcomes inside the window.

- **Assignment mechanism known in** W. This condition requires that the joint probability distribution of the scores in W be known. If this holds, the assignment mechanism for $\mathbf{T}_W$ will also be known, which is equivalent to the condition of known assignment in randomized experiments. The general idea is that the researcher assumes that all units with scores within the window W had an ex-ante probability of being assigned to treatment versus control given by a known mechanism.
- **Lack of relationship between score and outcomes in** W. This assumption requires that

the potential outcomes $(\mathbf{Y}_W(\mathbf{0}), \mathbf{Y}_W(\mathbf{1}), \mathbf{D}_W(\mathbf{0}), \mathbf{D}_W(\mathbf{1}))$ not be functions of $\mathbf{X}_W$. This requirement stems from the fundamental difference between this framework and an actual randomized experiment. Imagine an experiment where the score is a randomly generated number from the uniform distribution between 0 and 100, and a treatment is administered to patients whose number is above 50. This experiment has all the features of an RD design, but it also differs from it in a crucial way: the score variable, by virtue of being a randomly generated number, is unrelated to potential outcomes at each value of the score. In contrast, in the continuity-based RD design framework, it is common for the score to be related to the potential outcomes. For example, it is generally understood that patients with higher oncotype scores are systematically different from patients with lower oncotype scores.

In the continuity-based RD design, the fact that the score is related to the potential outcomes does not present any challenges because the parameter of interest is defined at the cutoff point. But in the local randomization approach, the parameter is defined in an interval W, and thus we must assume that the value of the score within this interval is unrelated to the potential outcomes—a condition that is guaranteed neither by the random assignment of the score X_i, nor by the random assignment of the treatment T_i. Such assumption is plausible for small neighborhoods around the cutoff, that is, for those units that have scores closest to the cutoff. This condition also implies the exclusion restriction we introduced in Section 3 requiring that T_i have no direct effect on Y_i.

The above discussion reveals that there are key differences between an actual randomized experiment with a randomly generated score, a continuity-based RD design, and a local randomization RD design. When the score is a generated random number, the potential outcomes are unrelated to the score for all possible score values, and there is no uncertainty about the functional forms of $\mathbb{E}[Y_i(1)|X_i = x]$ and $\mathbb{E}[Y_i(0)|X_i = x]$: they are constant functions of x. In contrast, in a continuity-based RD design, the potential outcomes can be related to the score everywhere; the functions $\mathbb{E}[Y_i(1)|X_i = x]$ and $\mathbb{E}[Y_i(0)|X_i = x]$ are unknown, and estimation and inference is based on approximating them at the cutoff. Finally, under the local randomization RD design, the potential outcomes can be related to the score far from the cutoff, but there is a window W around the cutoff where this relationship is assumed to vanish. More specifically, in the canonical local randomization

framework, the regression functions $\mathbb{E}[Y_i(1)|X_i = x]$ and $\mathbb{E}[Y_i(0)|X_i = x]$ are unknown over the entire support of the score variable, but inside W they are assumed to be constant functions of x.[2]

In the local randomization framework, we can define parameters of interest that are analogous to those discussed in the continuity-based framework. The main difference is that the continuity-based parameters are defined at the cutoff, and the analogous local randomization parameters are defined in the window W around the cutoff. The local randomization sharp RD parameter in the super-population setup is the average treatment effect inside the window W, analogous to $\tau_{\texttt{SRD}}$, defined as

$$\lambda_{\texttt{SRD}} \equiv \mathbb{E}[Y_i(1) - Y_i(0)|X_i \in W],$$

where, as before, the potential outcomes $Y_i(1), Y_i(0)$ are seen as independent and identically distributed (i.i.d.) draws from a super-population. Under the local randomization assumptions, $\lambda_{\texttt{SRD}} = \mathbb{E}[Y_i|T_i = 1, X_i \in W] - \mathbb{E}[Y_i|T_i = 0, X_i \in W]$, which is estimable from the data.

When there is non-compliance, the sharp RD estimator of the effect of T_i on Y_i and the effect of T_i on D_i estimate, respectively, the following parameters,

$$\lambda_{\texttt{Y}} \equiv \mathbb{E}[Y_i|T_i = 1, X_i \in W] - \mathbb{E}[Y_i|T_i = 0, X_i \in W]$$
$$\lambda_{\texttt{D}} \equiv \mathbb{E}[D_i|T_i = 1, X_i \in W] - \mathbb{E}[D_i|T_i = 0, X_i \in W],$$

which are analogous to the continuity-based parameters $\tau_{\texttt{Y}}$ and $\tau_{\texttt{D}}$. Under the local randomization assumptions, the parameters $\lambda_{\texttt{Y}}$ and $\lambda_{\texttt{D}}$ capture the average effect of assigning the treatment for observations with scores in the window, that is, $\lambda_{\texttt{D}} = \mathbb{E}[D_i(1) - D_i(0)|X_i \in W]$, and $\lambda_{\texttt{Y}} = \mathbb{E}[Y_i(1, D_i(1)) - Y_i(0, D_i(0))|X_i \in W]$.

Finally, we can also define the local-randomization fuzzy RD parameters as the ratio

$$\lambda_{\texttt{FRD}} \equiv \frac{\lambda_{\texttt{Y}}}{\lambda_{\texttt{D}}}.$$

As in the continuity-based framework, under appropriate assumptions, $\lambda_{\texttt{FRD}}$ can be interpreted as the average treatment effect in the window for compliers. The assumptions typically used are

[2] While allowing for dependence between potential outcomes and the score in the local randomization framework is possible by means of score-adjustments (Cattaneo et al., 2017), we postpone discussion of this generalization until we discuss estimation and inference methods.

similar to those required in IV settings, now applied to observations with scores in the window W, and are similar to those discussed for the continuity-based framework. The effect of the treatment assignment on the treatment received, $\lambda_{\mathtt{D}}$, must be well separated from zero. The exclusion restriction that the treatment assignment has 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 W. Finally, the assumption of monoticity requires that there are no units with scores in W who receive a treatment condition that is always opposite to their assignment.

4.3.1 Estimation and Inference: Local Randomization Methods

The practical implementation of the local randomization framework requires two steps: (i) choosing the window 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. Window selection is the most important step in the implementation of the local randomization approach for RD analysis. Although W could be selected in an ad-hoc fashion, the more principled approach is to select it using predetermined covariates—characteristics of the units whose values are determined before the treatment is assigned. The window selection process we discuss was first proposed in Cattaneo et al. (2015) and is now the standard method used in the empirical implementation of the RD local randomization framework.

The selection process is based on the idea of covariate balance in randomized controlled trials. In a randomized experiment, successful randomization of the treatment implies that the distribution of pre-intervention covariates is equal between control and treatment groups. Consequently, if the local randomization assumptions hold in a window W, we should find that predetermined covariates are balanced for units whose scores lie inside W. Building on this observation, the window selection procedure chooses the largest window around the cutoff such that covariate balance holds in that window and in all windows contained in it. In other words, W is chosen to be 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 sub-windows contained in it.

This window selection method requires that there be at least one important predetermined

covariate, Z, that is related to the score variable everywhere except inside W. Typically, Z will contain a number of variables. Once these predetermined covariates are chosen, the implementation of the window selection can be based on Fisherian or super-population methods. However, it is often better to employ Fisherian methods because the smaller windows may have very few observations, which can invalidate the use of large-sample approximations. For implementation, first a test statistic and an assignment mechanism are chosen, and then the researcher performs a sequence of hypothesis tests that test the sharp null hypothesis that the treatment has no effect on the covariate 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 (i.e., 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.

In Section 4.5, we use the ART application to illustrate this window selection method as well the other methods of analysis for RD designs under local randomization.

Estimation and inference inside the window. The most common approach for the analysis and interpretation of experiments and observational studies in the social, behavioral, and biomedical sciences is the super-population setup. In this large-sample approach, the potential outcomes are considered to be random variables, conceptually understood as a sample from a hypothetical super-population, and thus estimation and inference reflects uncertainty from two sources simultaneously: treatment assignment mechanism and random sampling (from the super-population). In the context of RD designs under a local randomization framework, this approach justifies using standard estimation and inference methods such as difference-in-means, linear regression methods, and IV methods, for those units whose scores lie within W. In other words, this approach reduces the RD analysis toolkit to methods based on large-sample approximations for observational or experimental data within the local randomization window. Because the super-population approach assumes that the underlying population is large within W (i.e., N_W is large), it is most appropriate when the sample size inside the window is sufficiently large.

We first focus on estimation and inference for the sharp RD parameter, $\lambda_{\texttt{SRD}}$. A natural estimator for this parameter is the difference in means between the observed outcomes in the treated and control groups, that is,

$$\widehat{\lambda}_{\texttt{SRD}} = \frac{1}{N_W^+} \sum_{i:X_i \in W} T_i Y_i - \frac{1}{N_W^-} \sum_{i:X_i \in W} (1 - T_i) Y_i, \tag{2}$$

where N_W^+ is the total number of units with $c + w \geq X_i \geq c$ (above the cutoff and inside W), N_W^- is the total number of units with $c > X_i \geq c - w$ (below the cutoff and inside W), and $N_W = N_W^+ + N_W^-$ is the total number of units inside W. Under the local randomization assumptions (plus perfect compliance), $\widehat{\lambda}_{\texttt{SRD}}$ is a consistent estimator of $\lambda_{\texttt{SRD}}$ provided that $N_W^- \to \infty$ and $N_W^+ \to \infty$. Furthermore, under the assumption of complete or fixed-margins randomization, the estimator is also unbiased.

When compliance is imperfect and the RD design is fuzzy, however, $\widehat{\lambda}_{\texttt{SRD}}$ no longer estimates the average effect of the treatment received at the cutoff. In this case, as in the continuity-based framework discussed in the prior section, researchers typically focus on the ITT effects of the treatment assignment, or on the effects of the treatment received for a subpopulation.

We can consistently estimate the sharp RD effects of T_i on Y_i and D_i, $\lambda_{\texttt{Y}}$ and $\lambda_{\texttt{D}}$, with the difference in the average observed outcomes between the treated and control groups inside the window,

$$\begin{aligned}
\widehat{\lambda}_{\texttt{Y}} &= \frac{1}{N_W^+} \sum_{i:X_i \in W} T_i Y_i - \frac{1}{N_W^-} \sum_{i:X_i \in W} (1 - T_i) Y_i, \\
\widehat{\lambda}_{\texttt{D}} &= \frac{1}{N_W^+} \sum_{i:X_i \in W} T_i D_i - \frac{1}{N_W^-} \sum_{i:X_i \in W} (1 - T_i) D_i.
\end{aligned}$$

The fuzzy RD estimator $\widehat{\lambda}_{\texttt{Y}}$ is identical to the sharp RD estimator: both are the difference in means between treated and control outcomes inside the window. The differences is that, in the case of noncompliance, the expectation or limit of this difference-in-means is no longer the average treatment effect, but rather the average effect of assigning the treatment. Similarly, the first stage estimand, which captures the average effect of the treatment assignment on the treatment take-up, is estimated with the difference in the share of treated units between both groups. Finally, we can

consistently estimate the local randomization fuzzy RD parameter, $\lambda_{\texttt{FRD}}$, with

$$\widehat{\lambda}_{\texttt{FRD}} = \frac{\widehat{\lambda}_{\texttt{Y}}}{\widehat{\lambda}_{\texttt{D}}}.$$

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 W is assumed to be large enough for distributional approximations to hold. The super-population 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, standardized least-squares and two-stage least-squares coefficients, etc., frequently used in the analysis of experiments.

Large-sample approximations can be a limitation for local randomization RD analysis, where it is common to have few observations with scores inside the window W where the two local randomization conditions are assumed to hold. When the number of observations in W is small, adopting a Fisherian setup may be more appropriate. In a Fisherian analysis, the data is seen as a fixed population instead of a random sample, and the emphasis is on inference rather than point estimation. The potential outcomes are seen as fixed quantities, and average effects are not well-defined; instead, researchers focus on the individual treatment effect $Y_i(1) - Y_i(0)$ and make inferences under the sharp null hypothesis that this effect is zero for all units. The advantage of the Fisherian framework is that, under a known assignment mechanism and the sharp null hypothesis, inferences can be conducted via permutation, and large-sample distributional approximations are not required; this approach is finite sample valid because, under the sharp null hypothesis, the distribution of any test statistic based on the data is exclusively governed by the treatment assignment mechanism. The Fisherian setup is most appropriate to conduct inferences when the number of observations is small and the randomization mechanism that assigned units to treated and control is either known or can be approximated.

In real experiments, the implementation of Fisherian inference is straightforward because the treatment assignment mechanism is known. A commonly used assignment is a fixed margins or complete randomization mechanism, where the number of treated and control units are fixed, and all assignments with the same numbers of treated and control units are equally likely. In addition to knowledge of the assignment mechanism, the implementation requires the choice of a test statistic.

Most applications employ the difference in means between control and treatment units, but other good choices are Kolmogorov-Smirnov (KS) statistic and the Wilcoxon rank sum statistic. See Ernst (2004) for an overview of permutation-based methods in statistics.

The same generic randomization inference ideas can be deployed to the specific context of RD designs, but only for units with score within the window W around the cutoff. In practice, two issues must be addressed. First, the window W is unknown and its half length plays the same role as a bandwidth in the continuity framework. As discussed above, this window should be chosen with principled and data-driven selection methods. Second, in contrast to real experiments, in the local randomization RD framework the assignment mechanism is typically unknown and thus must be assumed. A natural strategy is to adopt a fixed-margins mechanism within W, assuming that the number of units above and below the cutoff is fixed. The fixed-margins assignment assumption is a natural choice in the RD context, but it is made for simplicity only; it can be substituted by other assignment mechanisms if the researcher has relevant information or wishes to assess the robustness of the results.

A Fisherian test of the sharp null hypothesis of no treatment effect for any unit within W will control Type I error for any sample size, and a p-value based on permuting the treatment assignment within W according to the pre-specified treatment assignment mechanism is readily available to assess the statistical significance of this null hypothesis. 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. A natural approach is given by Imbens and Rosenbaum (2005), who also point out that Fisherian IV methods provide correct inferences even when the instrument is weak (i.e., when $\lambda_{\mathtt{D}}$ is close to zero). Their approach is not based on the standard two-stage least squares (2SLS) point estimator, but provides confidence intervals for $\lambda_{\mathtt{FLR}}$ under specific treatment effect modelling assumptions. See also Keele et al. (2017), Kang et al. (2018), and references therein, for recent methodological developments on IV methods, and Cattaneo et al. (2015), Cattaneo et al. (2017), and Cattaneo et al. (2022) for more details on local randomization RD analysis.

4.4 Evaluating the RD Assumptions

While RD designs are frequently described as a type of natural experiment (Titiunik, 2021), the researcher does not precisely control treatment assignment as would be the case in a randomized experiment. Thus, it is important to remember that RD designs remain a type of observational study and their key underlying assumptions are not guaranteed to hold by design (Sekhon and Titiunik, 2016, 2017).

The main threat to the validity of any RD design is the possibility that units are able to affect their scores to select their preferred treatment condition. If the cutoff determining treatment assignment is known to the units, and they have the ability of changing or "manipulating" their score, units may systematically select into any treatment condition of their choosing. Such selection process usually invalidates the underlying identifying assumptions of the RD framework. For example, from a continuity perspective, systematic placements of units above and below of the cutoff will lead to a discontinuity in their potential outcomes regression functions. Similar issues will arise in a local randomization framework, as units below and above the cutoff will be different for reasons other than their treatment assignment whenever they can systematically manipulate their score assignment.

Given these concerns, it is important to provide qualitative information about the administrative process by which scores are assigned and cutoffs are determined (including whether this information is public knowledge), and about how treatment assignment actually occurred in each application. 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, we would naturally expect that patients might try to manipulate their CD4 count in order to qualify for ART. However, so long as the CD4 count is determined by laboratory procedures that cannot be influenced by patients or physicians, this is less of a concern.

When compared to other observational studies, one strength of the RD design is that its key underlying assumptions often have several "empirical implications" that should hold and can provide indirect quantitative evidence about the design's validity. As such, all empirical studies based on a RD design should always include a full set 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 causal effects in addition to predicting the presence of an effect.

We review four key falsification and diagnostic tests for RD designs: (i) studying the density of the score near the cutoff, (ii) testing for effects on pre-treatment covariates or placebo outcomes, (iii) assessing the sensitivity of estimates to nearby to the cutoff units, and (iv) estimating treatment effects at alternative non-cutoff values of the score. All these methods are about general features of the RD design, and thus applicable in both sharp and fuzzy RD settings. In addition, similarly to IV settings, we stress the importance of checking the strength of the first-stage estimate in the fuzzy RD design (i.e., $\tau_{\mathtt{D}}$ and $\lambda_{\mathtt{D}}$ should be well-separated from zero).

4.4.1 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 considerably different from the number of observations above it (McCrary, 2008). 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. In other words, there is usually no reason for having a disproportionately larger number of units in one group relative to the other group near the RD cutoff point, beyond what would be generated by the specific treatment assignment mechanism. 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.

- *Binomial Test.* This approach was introduced by Cattaneo et al. (2015, 2017), building on the local randomization framework. The method begins by postulating a probability of treatment assignment for units in the window W around the cutoff, which in the context of RD designs is equivalent to a probability of receiving a score above the cutoff. The default choice is 50%—equal probability of being placed below or above the cutoff. Then, by virtue of random sampling, the number of control and treated units has an exact, finite sample binomial distribution in

W. Consequently, a binomial test is applicable: this hypothesis test assesses the likelihood of the observed configuration of control and treatment units, given a hypothesized probability of treatment assignment. Large differences between the number of control and treatment units within W lead to rejection of the null hypothesis, and thus is evidence against the RD design. In practice, this is formalized by reporting a p-value from a sequence of binomial tests for different windows around cutoff.

- *Density Test.* This approach was introduced by McCrary (2008), and is based on the continuity framework for RD analysis. The core idea is that, in the absence of manipulation, the density of the score variable should be continuous at the cutoff. This test can be conducted informally by plotting a histogram of the score, and examining it for "bunching" at or near the cutoff. In fact, such plot is a graphical representation of the binomial test. A more formal test is available using local polynomial approximations of the underlying density function of the score variable. Cattaneo et al. (2020b) developed a local polynomial density estimator that does not require pre-binning of the data, and exhibits size and power improvements relative to other approaches available in the literature. Formally, the null hypothesis is that the score density is continuous at the cutoff, and the nonparametric testing procedure implements a local polynomial estimator of the density of observations near the cutoff, separately for observations above and below the cutoff. A large sample approximation is used to tabulate a rejection rule, based on MSE-optimal bandwidth selection and robust bias correction inference methods analogous to those described in Section 4.2.1 for local polynomial RD inference.

4.4.2 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 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.

The implementation of this diagnostic can be done using both the continuity and local randomization frameworks. The underlying assumptions and methods for each case are analogous to those described in Sections 4.2.1 and 4.3.1, respectively, with the only change that now the outcome variable is either a predetermined covariate or a placebo outcome. In the case of continuity-based methods, because the regression functions of interest change when the outcome variable changes, it is important to re-estimate the bandwidth used for estimation and inference. This follows from the fact that the optimal bandwidths discussed in Section 4.2.1 trade off bias and variance, either in a mean squared error sense or in a coverage error sense, and both change when the regression function of interest changes. It is therefore incorrect to employ the same MSE-optimal or CE-optimal bandwidth for both point estimation of the main RD effect of interest and for falsification purposes, the same way that it would be incorrect to use the same bandwidth to analyze multiple outcome variables of interest.

In the case of the local randomization framework, some predetermined covariates Z will be used to select the window. If the researcher is in fact able to find a window W of non-zero length where these Z covariates are balanced, this simultaneously provides a falsification of the RD assumptions based on Z. In contrast, if there is no window of positive length where these covariates are balanced, this can also be interpreted as empirical evidence against the assumption of local randomization. Unlike the case of continuity-based methods, in this setting it is both appropriate and advisable to employ the same window W for falsification testing as well as for estimation and inference of RD treatment effects for multiple outcome variables. Conceptually, this method imposed substantially stronger assumptions on the underlying distribution of the data, thereby treating the RD design as a local randomized experiment for units with score within W and therefore justifies focusing on this subsample of the data for the entire analysis.

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.

4.4.3 Sensitivity to Bandwidth or Window Selection

The choice of window or bandwidth where estimation and inference are conducted are extremely important in any RD analysis. Because results are typically altered by changes to these choices, it can be important to study whether the conclusions of the analysis are robust to different choices of bandwidth or local randomization window. To implement this sensitivity analysis, researchers can re-estimate the RD treatment effect for bandwidths or windows that are smaller or larger than the one originally chosen. In the continuity-based framework, if the original bandwidth is MSE-optimal, considering larger bandwidths is not advisable because of the potential for misspecification bias. In the local randomization framework, considering larger windows is not advisable if important covariates become imbalanced when larger windows are considered.

4.4.4 Donut Hole

This diagnostic is based on the idea that the few observations closest to the cutoff should not drastically determine the empirical results. Polynomial approximations in RD designs could suffer from biases near the cutoff because of Runge's phenomenon. As a response to this concern, the so-called "donut hole" falsification test removes a few observations closest to the cutoff in an attempt to understand the sensitivity of the results to those observations. In practice, this method is easily implemented by using either the continuity 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. In other words, a sequence of increasing, nested "holes" around the cutoff are generated, which remove the observations with scores lying on that region, and then the estimation and inference methods are re-implemented on the remaining sample.

This method is conceptually straightforward, and easy to implement in practice. An important caveat is that, unlike the case of predetermined covariates and placebo outcomes, in this case the continuity framework based on local polynomial estimation and inference should retain the same bandwidth as originally used, instead of re-estimating a new bandwidth for each new subsample generated by the donut hole. The reason is that here the goal is to understand sensitivity to the fit over the bandwidth used, and thus changing the bandwidth along the way would tamper the

analysis.

In applications, this falsification test is usually presented in a plot where the x-axis marks the different endpoints of the donut holes considered on the score dimension, and the y-axis reports the new point estimator and confidence interval obtained from the corresponding subsample of observations. If the resulting point estimates are roughly stable across the different donut holes considered, this provides empirical evidence of robustness. See Cattaneo et al. (2020a, Section 5) for more discussion on this method.

4.4.5 Placebo Cutoffs

This is the final falsification test we discuss for generic RD designs. The idea underlying this test 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 test analyzes units below and above the cutoff separately, generating a sequence of placebo or artificial RD cutoffs to check that there is no RD treatment effect at those alternative cutoffs. Empirical evidence of treatment effects at the placebo or artificial cutoffs may undermine the design if the researcher cannot explain why these effects occur. In this case, non-zero effects at placebo cutoffs suggest the possibility of other factors affecting the units of analysis in the background.

In the continuity-based framework, this test becomes a test of continuity of the regression functions away from the cutoff, which ideally should not lead to rejection of the null hypothesis of continuity because otherwise such rejection would suggest misspecification of the local polynomial approximation at the very least. More importantly, such rejection might be interpreted as evidence against the required continuity assumptions at the cutoff needed for identification, estimation, and inference. Similarly, in the case of the local randomization framework, this test is implemented in the same way as in the main analysis, but replacing the true cutoff with an artificial cutoff. In this context, this falsification analysis is a test of the functional form of the potential outcomes in relation to the score, with a non-zero effect at a placebo cutoff showing a non-constant relationship.

Regardless of the methods used, and of their interpretation, this falsification test is similar in spirit to the one based on predetermined covariates and placebo outcomes. In both cases, the ultimate goal is to generate new RD designs where null treatment effect is expected. In practice, this test is usually presented by means of a plot, as in the donut hole test, where the x-axis is

centered at the true cutoff and other artificial cutoffs are reported, both to the left and to the right, and the y-axis reports point estimates and confidence intervals at each cutoff presented. Of course, for each artificial cutoff, only observations below or above the true cutoff are used, as appropriate, to avoid contaminating the falsification analysis with the potentially non-zero true treatment effect. The expectation is that a treatment effect should occur only at the true cutoff and not at other artificial cutoffs, where treatment status is constant by construction. A different method to generate random, placebo cutoffs was proposed by Ganong and Jäger (2018). See Cattaneo et al. (2020a, Section 5) for more discussion on this method.

4.4.6 Fuzzy RD Validation

To close the discussion of falsification methods for RD designs with continuous score, we discuss some additional empirical validation methods that are specific for fuzzy RD designs. These methods are not necessarily about the validity of the RD design itself, but rather about the validity of the estimation and inference methods used for fuzzy RD designs. Since the fuzzy RD design shares features of both RD and IV designs, most of the methods used for estimation and inference in such setting require additional assumptions. The plausibility of some of these assumptions can be explored empirically. See Glymour et al. (2012), Baiocchi et al. (2014), Imbens and Rubin (2015), Pizer (2016), Keele et al. (2019), 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. Depending on the specific RD framework considered, a parameter of interest in the fuzzy RD setup is $\tau_{\texttt{FRD}} = \tau_{\texttt{Y}}/\tau_{\texttt{D}}$ or $\lambda_{\texttt{FRD}} = \lambda_{\texttt{Y}}/\lambda_{\texttt{D}}$, with their respective estimators $\widehat{\tau}_{\texttt{FRD}} = \widehat{\tau}_{\texttt{Y}}/\widehat{\tau}_{\texttt{D}}$ and $\widehat{\lambda}_{\texttt{FRD}} = \widehat{\lambda}_{\texttt{Y}}/\widehat{\lambda}_{\texttt{D}}$. This implies that $\tau_{\texttt{D}}$ and $\lambda_{\texttt{D}}$ should be far enough from zero and, by implication, so should $\widehat{\tau}_{\texttt{D}}$ and $\widehat{\lambda}_{\texttt{D}}$ in finite samples. 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 (Feir et al., 2016). A weak IV test examines whether the effect of the instrument on the treatment exposure is sufficiently strong—typically based on F-tests. 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 neighborhood around the cutoff—determined by either a bandwidth in the continuity framework, or a randomization window

in the local randomization framework. 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. Finally, we note that some inference methods are robust to the problem of weak-identification, such as the Fisherian methods discussed in the context of local randomization RD analysis.

Similarly to the sharp RD design, predetermined covariates can 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 cutoff. In the continuity-based framework, this is done by employing local polynomial methods and treating each covariate as an outcome in a standard sharp RD analysis. In the local randomization framework, the chosen inference setup is used to test whether each covariate is balanced in the chosen window around the cutoff. In both cases, the balance tests use only the predetermined covariates and the score. The treatment received is not used for covariate falsification purposes.

Another popular approach in IV biomedical research is to report 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. (2017), Zhao and Small (2018), and references therein, for related formal and graphical methods, which do not discuss here to conserve space.

The key exclusion restriction in RD and IV settings—that being assigned to the treatment has no effect on the outcome except via actual treatment exposure—is untestable, just like the continuity and local randomization RD assumptions. These are fundamental identifying assumptions. For fuzzy RD designs, it is important to evaluate the exclusion restriction using qualitative information. In addition, researchers can use some falsification tests that may offer empirical evidence in support of the exclusion restriction. For example, a test of this type would identify a subgroup where all the units are exposed to the treatment, and then estimate whether the treatment assignment (i.e., the instrument) has any effect on the outcome in this always-treated subgroup of patients. If the exclusion restriction holds, the instrument should have no effect on the outcome within this subgroup.

4.5 Empirical Illustration

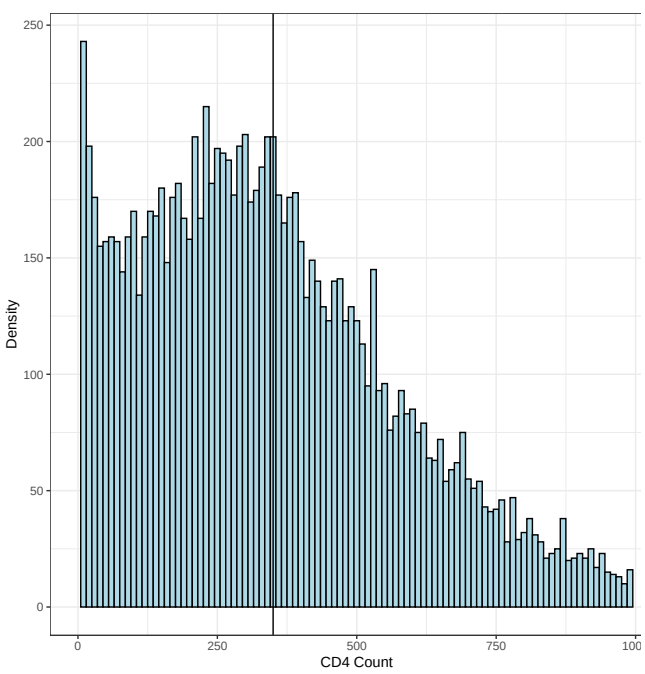
We illustrate empirically all the ideas and concepts discussed so far using the ART application. We first focus on results employing continuity-based methods, and then present an analysis based on the local-randomization framework. This application allows us to use both the continuity-based and the local randomization frameworks for RD analysis, because the CD4 count is a running variable with a large number of unique values. We have $11,306$ observations in our data, of which $1,229$ have a unique CD4 count. We start by treating this running variable as approximately continuous and employing continuity-based methods. We then use local randomization methods, which do not require the score to be continuous. In each case, we proceed in two phases. First, we conduct falsification tests that focus on validating the RD assumptions invoked. Then, we report the main results followed by some tests to asses the robustness of the main conclusions.

4.5.1 Continuity-Based Methods

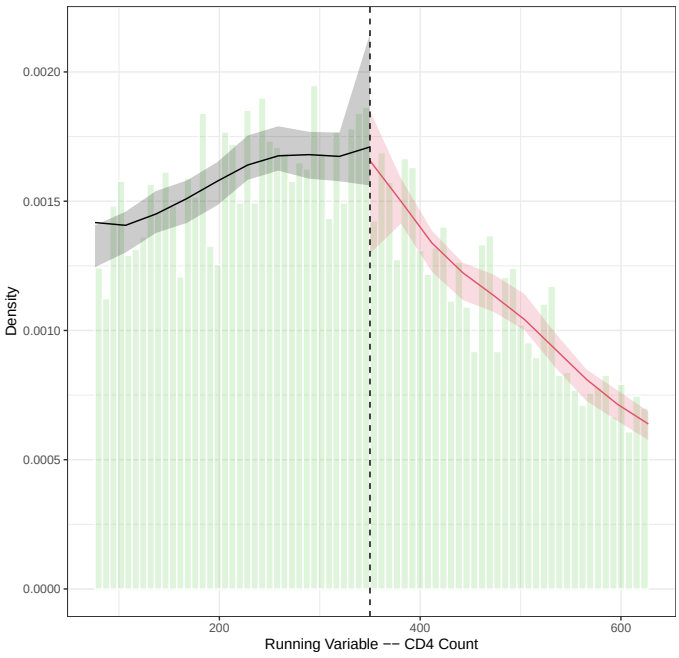
We first focus on the falsification test based on the density of the score around the cutoff. Figure 2a contains the raw histogram of the score in this application, the CD4 count. Informally, there appear to be no obvious signs of bunching near the 350 cutoff. We also implement the formal density test based on local polynomials, which we illustrate in Figure 2b. 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 ($p = 0.1858$). We also reject the hypothesis of a change in density near the cutoff using a binomial test in the window $[349, 351]$ around the 350 cutoff ($p = 0.2026$). 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.

Next, we estimate the RD treatment effect for predetermined covariates. Table 1 analyzes the available baseline covariates in the data. The results are calculated by using robust local polynomial methods to estimate the RD effect for each of the baseline covariates, treating each 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.

Figure 2: Density of the RD score around the cutoff for ART application.



(a) Histogram of CD4 count with threshold marked.



(b) Histogram of CD4 count with adaptive local polynomial estimator and 95% confidence intervals.

Next, we focus on two tests that are specific to RD designs with non-compliance: 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 507, well above the standard critical value thresholds used in the IV literature.

We also test for differences in baseline covariates at the cutoff using the fuzzy RD estimator. The results are in Table 2, and they are similar to those results in Table 1 based on the intention-to-treat RD estimator: we still find that there are no significant differences in the baseline covariates at the cutoff.

We now turn to the effects on the main outcome of interest, reported in Table 3. First, we focus on the effect of being assigned to treatment (in this case, having a score below the cutoff).

Table 1: Intention-to-treat RD Effects of ART initiation on Predetermined Covariates

	Mean Below	Mean Above	$\widehat{\tau}_{\mathtt{Y}}$	Robust p-value	CER-Optimal Bandwidth	N_h^-	N_h^+
Age 0-18	0.06	0.07	0.02	0.41	78.88	1453	1265
Age 18 -25	0.27	0.29	0.02	0.66	82.53	1540	1325
Age 25-30	0.24	0.21	-0.03	0.41	95.69	1782	1485
Age 30-35	0.13	0.13	0.00	0.96	70.83	1330	1152
Age 35-40	0.09	0.09	-0.00	0.93	86.83	1613	1372
Age 40-45	0.07	0.08	0.01	0.70	66.55	1251	1090
Age 45-55	0.10	0.10	-0.00	0.93	83.68	1561	1337
Age 55+	0.04	0.03	-0.01	0.41	59.19	1119	998
2011 Qtr3	0.13	0.11	-0.02	0.46	89.89	1712	1441
2011 Qtr4	0.20	0.17	-0.03	0.42	70.77	1357	1175
2012 Qtr1	0.19	0.21	0.02	0.45	95.57	1819	1512
2012 Qtr2	0.18	0.19	0.00	0.91	69.91	1339	1155
2012 Qtr3	0.17	0.19	0.02	0.57	85.46	1624	1383
2012 Qtr4	0.14	0.13	-0.01	0.75	99.04	1891	1569
Female	0.66	0.73	0.07	0.11	56.20	1069	964
Clinic A	0.15	0.12	-0.03	0.29	60.26	1160	1026
Clinic B	0.13	0.15	0.02	0.46	67.91	1297	1121
Clinic C	0.14	0.15	0.02	0.56	74.34	1416	1229

Note: Each row reports the average effect (at the cutoff) of being assigned to the treatment versus the control condition on a given predetermined covariate. Analysis based on local linear estimation with coverage-error (CE) 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, $\widehat{\tau}_{\mathtt{Y}}$, 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 CE-optimal bandwidth, and the last two columns report the number of observations below and above this bandwidth, respectively.

We find that having a CD4 count of 350 or greater reduces the likelihood of ART initiation by 20 percentage points ($\widehat{\tau}_{\mathtt{D}}$) and also reduces program retention by 10 percentage points ($\widehat{\tau}_{\mathtt{Y}}$). 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, $\widehat{\tau}_{\mathtt{FRD}}$. We find that ART initiation increases program retention by 60 percentage points, and the confidence interval is bounded away from zero. Thus, we find 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 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 now probe the robustness of the results in Table 3. First, we estimate the RD effect using two placebo cutoffs: 300 and 400. Table 4 contains the results. For the intention-to-treat effect ($\widehat{\tau}_{\mathtt{Y}}$) on program retention, we find that that the robust 95% confidence interval barely includes

Table 2: Fuzzy RD Effects of ART initiation on Predetermined Covariates

	$\widehat{\tau}_{\texttt{FRD}}$	Robust p-value	CER-Optimal Bandwidth	N_h^-	N_h^+
Age 0-18	-0.08	0.39	78.01	1453	1265
Age 18 -25	-0.02	0.92	63.27	1191	1041
Age 25-30	0.10	0.65	68.61	1293	1113
Age 30-35	-0.01	0.97	63.16	1191	1041
Age 35-40	0.03	0.83	67.55	1270	1099
Age 40-45	-0.06	0.66	54.63	1020	928
Age 45-55	0.01	0.94	67.23	1270	1099
Age 55+	0.06	0.43	72.88	1357	1170
2011 Qtr3	0.09	0.54	70.25	1357	1175
2011 Qtr4	0.41	0.15	49.65	936	870
2012 Qtr1	-0.12	0.52	69.08	1339	1155
2012 Qtr2	-0.10	0.67	54.15	1041	948
2012 Qtr3	-0.12	0.52	69.86	1339	1155
2012 Qtr4	-0.05	0.83	61.77	1176	1040
Female	-0.60	0.13	47.61	879	828
Clinic A	0.18	0.37	48.01	912	857
Clinic B	-0.11	0.55	55.63	1060	961
Clinic C	-0.08	0.58	70.38	1357	1175

Note: The first column is the fuzzy RD effect ($\widehat{\tau}_{\texttt{FRD}}$) for each predetermined covariate. Analysis based on local linear estimation with coverage-error (CE) optimal bandwidth. P-value based on robust bias correction inference methods. The third column reports the CE-optimal bandwidth, and the last two columns report the number of observations below and above this bandwidth, respectively.

zero. However, the estimated fuzzy RD effect of ART uptake on program retention is significantly different from zero, with robust 95% confidence interval of $[0.21, 0.98]$ that shows a positive effect of having a CD4 count below 350 on program retention.

Finally, we report the results from the donut hole diagnostic test. For this test, we dropped patients with CD4 count values of 349, 350, and 351 and re-estimated the RD effects. We report the results in Table 5. We find there are 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.

As we noted earlier, a patients' CD4 count is actually discrete, but we deemed it close enough to continuous to use continuity-based RD methods. In the upcoming section, we use local randomization methods to account for the discrete nature of the score directly.

4.5.2 Local Randomization Methods

Under the local randomization approach, the analysis begins with window selection. This window is chosen by selecting the largest region around the cutoff where baseline covariates are balanced.

Table 3: RD Estimates of ART Assignment and Initiation

	RD Effect	95% Robust CI	Bandwidth (h)	N_h^-	N_h^+
ITT Effect of ART Assignment on ART Initiation	-0.21	[-0.28,-0.11]	110.62	1446	1147
ITT Effect of ART Assignment on Program Retention	-0.14	[-0.22,-0.04]	110.62	1446	1147
Fuzzy Effect of ART Initiation on Program Retention	0.66	[0.32,1]	110.62	1446	1147

Note: The rows show, respectively, $\widehat{\tau}_{\mathtt{D}}$, $\widehat{\tau}_{\mathtt{Y}}$, and $\widehat{\tau}_{\mathtt{FRD}}$. Analysis based on local linear estimation with mean-squared-error (MSE) optimal main bandwidth, h, reported in third column. Column labeled "95% Robust CI" reports the robust 95% confidence intervals based on robust bias-corrected inference. The last two columns report the number of observations below and above bandwidth h, respectively. Bias bandwidth used is 204.61.

Table 4: RD Estimates of ART Assignment – Placebo Cutoffs of 300 and 400

		RD effect	95% Robust CI	Bandwidth (h)	N_h^-	N_h^+
$c = 300$	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
$c = 400$	ITT Effect of ART Assignment on ART Initiation	-0.01	[-0.11,0.04]	45.42	785	642
	ITT Effect of ART Assignment on Program Retention	0.004	[-0.17,0.15]	55.51	602	508

Note: The rows show, respectively, $\widehat{\tau}_{\mathtt{D}}$, $\widehat{\tau}_{\mathtt{Y}}$, and $\widehat{\tau}_{\mathtt{FRD}}$. Analysis based on local linear estimation with mean-squared-error (MSE) optimal main bandwidth, h, reported in third column. Column labeled "95% Robust CI" reports the robust 95% confidence intervals based on robust bias-corrected inference. The last two columns report the number of observations below and above bandwidth h, respectively. Bias bandwidths used in rows one through four are 39.61, 47.14, 64.43, and 69.34, respectively.

Under local randomization, window selection is congruent with the RD falsification test based on predetermined covariates. If we cannot find any neighborhood around the cutoff where the covariates are balanced, that is evidence against the design. 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–a total of 176 patients. We omit the balance test results for space considerations, but the window selection was based on the same covariates shown in Table 1. This window is much narrower than the bandwidth estimated using continuity-based methods. Specifically, the estimated bandwidth in the continuity-based analysis for the fuzzy RD design shown in Table 3 was 88.6, and the analysis was based on the 2,117 observations with CD4 counts approximately between 261 and 439 (350 ± 89). This reduction in the number of observations from 2,117 in the continuity-based framework to 176 in the local randomization framework is typical: local randomization methods, by their very nature, focus a much smaller neighborhood around the cutoff.

We now report the main RD analysis under the local randomization framework. The results,

Table 5: RD Estimates of ART Assignment and Initiation – Donut Hole Falsification

	RD effect	95% Robust CI	Bandwidth (h)	N_h^-	N_h^+
ITT Effect of ART Assignment on ART Initiation	-0.23	[-0.3,-0.13]	118.12	1539	1220
ITT Effect of ART Assignment on Program Retention	-0.13	[-0.21,-0.03]	118.12	1539	1220
Fuzzy Effect of ART Initiation on Program Retention	0.58	[0.24,0.88]	118.12	1539	1220

Note: The rows show, respectively, $\widehat{\tau}_{\mathtt{D}}$, $\widehat{\tau}_{\mathtt{Y}}$, and $\widehat{\tau}_{\mathtt{FRD}}$. Analysis based on local linear estimation with mean-squared-error (MSE) optimal main bandwidth, h, reported in third column, but excluding observations with CD4 count equal to 349, 350, and 351. Column labeled "95% Robust CI" reports the robust 95% confidence intervals based on robust bias-corrected inference. The last two columns report the number of observations below and above bandwidth h, respectively. Bias bandwidth used is 203.81.

presented in Table 6, are very different from the continuity based results. Both the intention-to-treat estimate, $\widehat{\lambda}_{\mathtt{Y}}$, and the first-stage estimate, $\widehat{\lambda}_{\mathtt{D}}$, estimates have the same sign as $\widehat{\tau}_{\mathtt{Y}}$ and $\widehat{\tau}_{\mathtt{D}}$ but are smaller and not statistically significant. The local randomization fuzzy RD estimate, $\widehat{\lambda}_{\mathtt{FRD}}$, has the opposite sign as $\widehat{\tau}_{\mathtt{FRD}}$ and the confidence intervals are not well-behaved as they both exceed 1 and -1. Note that these confidence intervals are based on two-stage least squares estimates.

The results obtained using local randomization and continuity-based methods can be compared and contrasted in terms of the different local neighborhoods employed by the two methods. To explore whether the difference in window width drives the results, we implemented the local randomization estimator using a series of wider windows. Specifically, we used three wider windows of 340–360, 335–365, and 330–370. Table 6 also contains the fuzzy RD estimates from the three wider windows. For the first wider window, the results are closer to those that used continuity based methods, while for the other two windows the results are nearly identical to those from continuity based methods.

This application illustrates one common way how continuity-based and local randomization analysis can sometimes differ. The local randomization framework requires the assumption that the scores and outcome are unrelated in a window around the cutoff, which is often plausible only for the few units with scores closest to the cutoff. By implication, local randomization methods may have low power in some applications because the neighborhood used is purposely chosen to be small. On the other hand, the continuity-based results use more extrapolation, imposing smoothness assumptions to ensure that observations further away from the cutoff can be used to estimate the effect at the cutoff. This conceptual contrast in estimation and inference approaches

Table 6: RD Estimates of ART Assignment and Initiation– Local Randomization Approach

	Risk Difference	95% Confidence Interval	N_W^-	N_W^+
ITT Effect of ART Assignment on ART Initiation	0.02	[-0.15 , 0.18]	62	59
ITT Effect of ART Assignment on Program Retention	-0.03	[-0.2 , 0.14]	62	59
Fuzzy Effect of ART Initiation on Program Retention	-1.5	[-23.43 , 20.43]	62	59
Wider Windows				
340-360				
ITT Effect of ART Assignment on ART Initiation	-0.14	[-0.25 , -0.04]	144	128
ITT Effect of ART Assignment on Program Retention	-0.09	[-0.2 , 0.02]	144	128
Fuzzy Effect of ART Initiation on Program Retention	0.62	[-0.03 , 1.27]	144	128
335-365				
ITT Effect of ART Assignment on ART Initiation	-0.21	[-0.3 , -0.13]	212	191
ITT Effect of ART Assignment on Program Retention	-0.11	[-0.2 , -0.02]	212	191
Fuzzy Effect of ART Initiation on Program Retention	0.51	[0.15 , 0.87]	212	191
330-370				
ITT Effect of ART Assignment on ART Initiation	-0.25	[-0.33 , -0.18]	277	247
ITT Effect of ART Assignment on Program Retention	-0.13	[-0.21 , -0.06]	277	247
Fuzzy Effect of ART Initiation on Program Retention	0.53	[0.27 , 0.8]	277	247

Note: In each panel, the first rows show the estimated first-stage effect, $\widehat{\tau}_{\mathtt{D}}$, the second row shows the estimated ITT effect on the outcome, $\widehat{\tau}_{\mathtt{Y}}$, and the third row shows the esitmated fuzzy RD effect $\widehat{\tau}_{\mathtt{FRD}}$. The top panel uses window $[346, 354]$.

is reflected in the empirical results presented above.

5 RD Analysis when the Score is Discrete

Our discussion so far has focused on RD designs where the score variable is a continuous measure (i.e., where each unit receives a unique score value). We now focus on RD designs where the score variable is coarse or discrete—that is, settings where several (and possibly many) units share the same score value. Continuity of the score is necessary for the key identifying assumption in the continuity-based RD framework, where it is assumed that certain features of potential outcomes (such as conditional expectations) under treatment and control are also continuous at the cutoff. However, as discussed previously, the RD design has long been motivated by the idea that around the cutoff assignment to the treatment occurs in an as-if randomized fashion rather than by the idea of continuity at the cutoff, which is closely related to the local randomization framework. In contrast to the continuity-based RD approach, the local randomization RD framework is applicable to settings where the score exhibits repeated values or mass points in its distribution. The drawback

of the local randomization approach is that it relies on design assumptions that are stronger than the continuity-based conditions. We discuss how to employ both RD frameworks when the running variable exhibits mass points, focusing on the changes in interpretation and implementation that are required. To avoid repetition, we only elaborate on the new aspects of identification, estimation, and inference that are required to handle discrete scores, building on the material presented in the previous section. We illustrate the methods with the chemotherapy and the cost-sharing application.

5.1 Illustrating the Design with RD Plots

In Section 4.1, we discussed how to use RD plots to illustrate the RD design when the running variable is continuous. We now discuss how to modify the construction of RD plots when the running variable is discrete and the data contains multiple units with the same score value.

We defined an RD plot as a graphical illustration that superimposes two different relationships in the same figure: a global polynomial fit of the outcome on the score, and a scatter plot of outcome means against bins of the running variable. In the continuous score context, the researcher must decide how to select the number and the type of bins; we discussed several methods to choose bins optimally based on different criteria. In contrast, when the running variable is discrete and the number of unique values of the score is small or relatively small, the most natural way to present binned means is to simply calculate the average outcome for every unique value of the score. In this sense, the optimal bins when the running variable is discrete are the unique values that the running variable takes.

To be more precise, if the score variable takes the values $\{\mathsf{x}_{K_-}, \ldots, \mathsf{x}_{-2}, \mathsf{x}_{-1}, \mathsf{x}_c, \mathsf{x}_1, \mathsf{x}_2, \ldots, \mathsf{x}_{K_+}\}$, with K_- denoting the number of unique values below the cutoff and K_+ denoting the number of unique values above the cutoff, the binned means graph is created by plotting (i) the sample average of the outcome for each score value, $\{\bar{Y}_{K_-}, \ldots, \bar{Y}_{-2}, \bar{Y}_{-1}, \bar{Y}_c, \bar{Y}_1, \bar{Y}_2, \ldots, \bar{Y}_{K_+}\}$—where $\bar{Y}_j = \frac{1}{\#\{i:x_i=\mathsf{x}_j\}} \sum_{i:x_i=\mathsf{x}_j} Y_i$ for $j \in \{K_-, \ldots, -2, -1, c, 1, 2, \ldots, K_+\}$ and $\#\{A\}$ denotes the number of elements in the set A—against (ii) the unique score values $\{\mathsf{x}_{K_-}, \ldots, \mathsf{x}_{-2}, \mathsf{x}_{-1}, \mathsf{x}_c, \mathsf{x}_1, \mathsf{x}_2, \ldots, \mathsf{x}_{K_+}\}$. (If there are no observations with score exactly equal to the cutoff, the sample mean $\bar{Y}_c$ will not be in the plot.)

Adding a global polynomial fit is not recommended when the score is discrete, unless the

number of unique score values is sufficiently large. For cases with few or moderately few unique score values, the global fit will result in overfitting and will not provide meaningful information about the outcome-score relationship.

When researchers are working with a discrete score application where the number of unique values taken by the score is very large, the recommendations above can be sidestepped and the continuity-based graphical methods of Section 4.1 can be applied either to the raw data, or after collapsing the data to one observation per unique score value, that is, to the pairs $\{\mathsf{x}_{K_-}, \bar{Y}_{K_-}\}, \ldots \{\mathsf{x}_{-2}, \bar{Y}_{-2}\}, \{\mathsf{x}_{-1}, \bar{Y}_{-1}\}, \{c, \bar{Y}_c\}, \{\mathsf{x}_1, \bar{Y}_1\}, \{\mathsf{x}_2, \bar{Y}_2\}, \ldots, \{\mathsf{x}_{K_+}, \bar{Y}_{K_+}\}$. See Cattaneo et al. (2022) for more details on collapsing the data when the running variable is discrete.

We illustrate with the chemotherapy application, which has a running variable, the oncotype score, that is discrete. Figure 3 shows the plot of the breast cancer recurrence indicator (the outcome of interest) on the oncotype score. Since this score takes only a smaller number of values, we simply plot the proportion of patients with breast cancer recurrence for each one of these values.

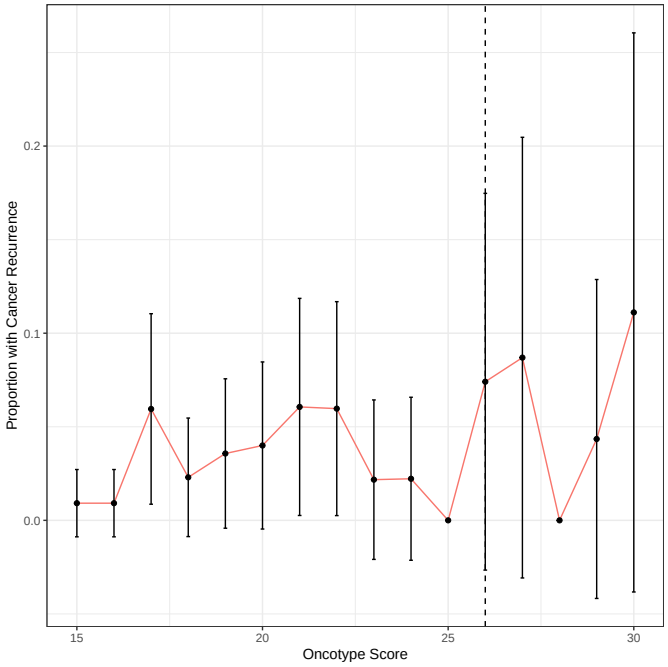


Figure 3: Proportion of patients with a recurrence of cancer by oncotype score. The x-axis is the discrete oncotype score.

5.2 Continuity-Based Methods

In RD settings where the score takes on a relatively large number of distinct values, even when there are repeated score values among the units, continuity-based methods can be applied if the researcher is willing to make additional assumptions. A fruitful approach is to view the number of unique values of the score 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 $N \leq n$ unique values among $X_1, X_2, \ldots, X_n$.

Whenever the number of the unique score values N is large enough, the continuity framework can be taken as valid for identification purposes. In other words, $\tau_{\texttt{SRD}}$, $\tau_{\texttt{Y}}$, $\tau_{\texttt{D}}$ and $\tau_{\texttt{FRD}}$ are reasonable quantities of interest. Intuitively, the key condition is whatever extrapolation takes place from the closest observed score value to the cutoff is sufficiently small. Clearly, some extrapolation would be needed, which is ultimately achieved via the estimation and inference methods employed, but with a large number N of unique score values, it is reasonable to rely on approximations that require minimal extrapolation to the cutoff.

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. Such extension is discussed in Calonico et al. (2019) and Calonico et al. (2020), and reviewed in Cattaneo et al. (2022). Heuristically, the only important change is related to the sample size used. 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 the number of unique values N—not the total number of observations, n. On the other hand, from an uncertainty perspective, either N 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 (i.e., $N < n$) affects bandwidth selection and inference methods, but the necessary modifications are straightforward and readily applicable in general purpose software. This logic justifies the empirical analysis in Section 4.5, where we used continuity-based methods despite having some repeated score values.

The situation is different when N is small, that is, when there is only a few unique values in

$X_1, X_2, \ldots, X_n$. 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 conditional expectation functions arbitrarily close to the cutoff. In this scenario, a reasonable 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. In this example, regardless of how large n is, we only have $N = 5$. 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, $\tau_{\mathtt{SRD}} = \mathbb{E}[Y_i(1)|X_i = c] - \mathbb{E}[Y_i(0)|X_i = c]$ 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.

The extreme example above shows a more general phenomenon: if the score is discrete, the canonical continuity-based RD parameters $\tau_{\mathtt{SRD}}$, $\tau_{\mathtt{Y}}$, $\tau_{\mathtt{D}}$ or $\tau_{\mathtt{FRD}}$ 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. The first approach restores the validity of continuity-based methods, while the second approach does not. In other words, continuity-based RD methods can be deployed to RD designs with discrete score variables whenever the local parametrizations are assumed to generate zero misspecification bias, that is, when the local polynomial model is assumed to be correctly specified. In this case, the total sample size n can be used for inference purposes, but bandwidth selection methods may still require sufficient variation (i.e., enough distinct values) in the score. For this reason, when N is very small, it is customary to choose a bandwidth by hand rather than in a data-driven fashion, circumventing bandwidth selection methods altogether. This is not a major concern, as many low-N applications have only a handful of possible values to consider.

If researchers are not willing to invoke stringent parametric assumptions to justify the use of continuity-based methods for RD analysis, they can choose the second approach and change the parameters of interest. In this case, local randomization methods can be applied with some modifications. We discuss this approach in the next section.

5.3 Local Randomization Methods

Provided the parameter of interest is changed or reinterpreted, 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 $K_- + K_+ + 1 = N$ the total number of unique values. In this context, the local randomization assumption reduces to specifying a window containing some of these unique values where the two local randomization conditions discussed above are assumed to hold.

We can define the following alternative RD parameters for settings where the score has few mass points:

$$\begin{aligned}
\theta_{\texttt{SRD}} &= \mathbb{E}[Y_i(1)|X_i = c] - \mathbb{E}[Y_i(0)|X_i = \mathsf{x}_{-1}], \\
\theta_{\texttt{Y}} &= \mathbb{E}[Y_i(1, D_i(1))|X_i = c] - \mathbb{E}[Y_i(0, D_i(0))|X_i = \mathsf{x}_{-1}], \\
\theta_{\texttt{D}} &= \mathbb{E}[D_i(1)|X_i = c] - \mathbb{E}[D_i(0)|X_i = \mathsf{x}_{-1}], \\
\theta_{\texttt{FRD}} &= \frac{\theta_{\texttt{Y}}}{\theta_{\texttt{D}}}.
\end{aligned}$$

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 only 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 W.

The choice of 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 W contains enough unique values of the score, it is also possible to also 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.

5.4 Evaluating the RD Assumptions

In Section 4.4, we discussed an array of falsification and validation methods for RD designs with a continuous score variable; these methods can be directly employed in settings where the score is discrete but the number of unique score values N is large enough. When N is small and there is only a few distinct score values in the data, some of these methods are easily applicable while others are not. We discuss specific details below.

- **Score Density near the Cutoff**. The binomial test continues to be valid regardless of the number of unique score values 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. First, if there is a mass point of observations at $X = c$ alone, this has to be removed before the analysis—such data configuration, for example, is common in some RD designs where the cutoff is determined by some rank of the score variable. Second, when N is large, the density test can still be used following the discussion above for local polynomial methods. Finally, when N is small, the test can only be used under the assumption that the polynomial approximation for the cumulative distribution function near the cutoff is correctly specified.
- **Predetermined Covariates and Placebo Outcomes**. Since this method is based on replacing an outcome variable of interest with either a predetermined covariate or a placebo outcome and repeating the main RD analysis, it remains applicable whenever the desired identification,

estimation, and inference methods remain valid. In other words, following the discussion in the previous sections, this method can be adapted or reinterpreted so that it remains valid in the case of a discrete score variable.

- **Sensitivity to Bandwidth or Window Selection**. Exploring the sensitivity of results to bandwidth selection will only be applicable when N is large enough and the researcher uses local polynomial methods. In that case, the same considerations apply as those discussed above for the case of continuous score. If researchers are using local randomization methods instead, the most natural sensitivity analysis is to analyze the effects in the smallest window, $[\mathsf{x}_{-1}, c]$, using Fisherian methods if the sample size is small. This recommendation does not apply if $[\mathsf{x}_{-1}, c]$ contains too few observations, as in this case even a Fisherian analysis will not be helpful because of possible lack of statistical power to detect effects.
- **Donut Hole**. This method is applicable whenever there are enough unique values in a neighborhood of the cutoff, so that some of the values closest to the cutoff can be removed and still leave enough observations for analysis. In cases where the number of unique values near the cutoff is very small, this validation analysis is harder to apply, as it requires assuming that the parametrization of the relationship between the potential outcomes and the score is correctly specified.
- **Alternative Cutoffs**. As in the case of the donut hole analysis, producing artificial cutoffs away from the real RD cutoff requires N to be sufficiently large. Since this method is implemented for units below and above the cutoff separately, it would not be applicable unless enough distinct score values are available on either side of the cutoff. When N is large enough, this method can be applied with the appropriate modifications needed to address the phenomenon of multiple units sharing the same score values.
- **Fuzzy RD Validation**. The validation analyses that we reviewed in Section 4.4 are adapted from the standard IV literature. Their applicability to the case of RD designs with discrete score depends on the specific method considered. In particular, methods for weak instruments and covariate balance can be adapted along the lines discussed above so they remain useful in a setting with discrete score.

5.5 Empirical Illustrations

5.5.1 Genetic Assay Guidelines for Chemotherapy

We first illustrate how to analyze an RD design with discrete score using the chemotherapy application. We begin by plotting the data. Figure 4 shows the proportion of patients who were given adjuvant chemotherapy by genetic score. Note that, given the presence of mass points, there is no need to optimally choose the number of bins to plot the data. Instead, we simply bin the data at each value of the score.

A few patterns emerge from Figure 4. First, this application is a clear fuzzy RD design where compliance with treatment is imperfect. While there is an increase in the proportion of patients that receive chemotherapy at the 26 cutoff, 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. Second, 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. In other words, the evidence 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.

Given that the number of unique values is $N = 16$, continuity-based methods are not really appropriate in this application. Instead, we choose to redefine the parameters of interest and apply local randomization methods. In particular, we focus on the parameters $\theta_{\texttt{SRD}}$, $\theta_{\texttt{Y}}$, $\theta_{\texttt{D}}$, and $\theta_{\texttt{FRD}}$, comparing observations with oncotype score equal to the "first" treated value of the score ($X_i = 26 = c$) against observations with oncotype score equal to the "last" control value:

$$\begin{aligned}
\theta_{\texttt{SRD}} &= \mathbb{E}[Y_i(1)|X_i = 26] - \mathbb{E}[Y_i(0)|X_i = 25], \\
\theta_{\texttt{IT}} &= \mathbb{E}[Y_i(1, D_i(1))|X_i = 26] - \mathbb{E}[Y_i(0, D_i(0))|X_i = 25], \text{ and} \\
\theta_{\texttt{FS}} &= \mathbb{E}[D_i(1)|X_i = 26] - \mathbb{E}[D_i(0)|X_i = 25].
\end{aligned}$$

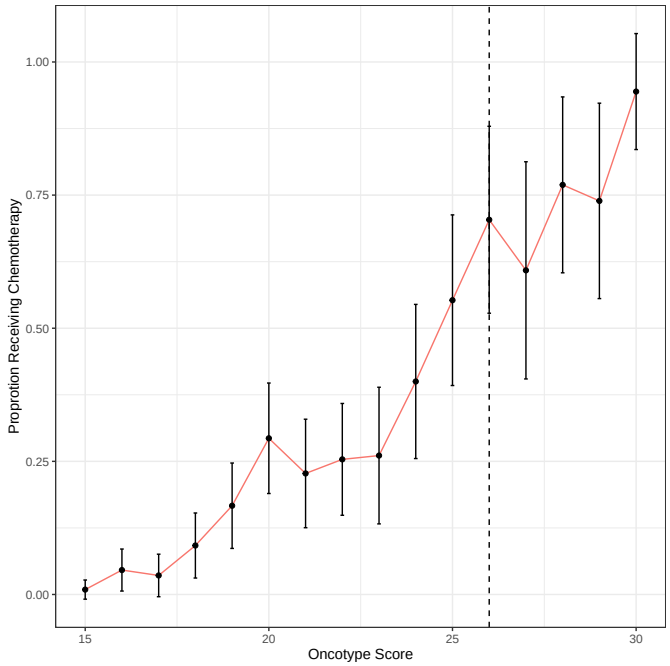


Figure 4: Proportion of patients receiving chemotherapy by oncotype score. The x-axis is the discrete oncotype score.

This redefinition of the parameters of interest is necessary because of the discreteness of the score, which prevents us from finding patients in the control group whose oncotype score is arbitrarily close to 26. Implicitly, our redefinition assumes that $\mathbb{E}[Y_i(0)|X_i = 25] = \mathbb{E}[Y_i(0)|X_i = 26]$, that is, that changing the oncotype score from 25 to 26 would not be enough to affect the average potential outcome in the absence of treatment. If this does not hold, $\theta_{\texttt{SRD}}$ will contain both the effect of the chemotherapy treatment and the effect caused by the one-unit difference in oncotype score. Analogous assumptions are implied for $\theta_{\texttt{Y}}$ and $\theta_{\texttt{D}}$.

We begin by validating the design, analyzing whether the number of treated and control observations is similar in a small neighborhood of the cutoff, and studying whether these observations are similar in terms of predetermined characteristics or covariates. 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 q is $1/2$, the p-value from the binomial test is 0.215. If instead we use a neighborhood of $W = [24, 27]$, 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 $1/2$. However, this may be an artifact

of the score distribution in this particular case. As oncotype scores increase, they become less common, and as such there is a clear downward trend in the density of the score. This suggests that, in this case, rejection of the null hypothesis might not necessarily be a sign of sorting.

We continue by exploring the local randomization window $W = [25, 26]$ using covariates. Given that we have few mass points, instead of using covariates to choose the window, we simply start with the smallest possible window, $[25, 26]$, and use covariates to validate it. Since most of the covariates are binary, we use the difference-in-means statistic to test for balance. Table 7 contains the estimated differences in means and p-values for a test of a difference in means. Note that we use Fisherian inference for these tests, since the sample sizes are relatively small. We find that the p-value for one covariate, tumor size, is less than 0.05. Does this imbalance imply that we should not proceed using an RD design under the local randomization framework? On the one hand, we can argue that only one of 16 covariates has a difference below 0.15. On the other hand, 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.

Table 7: Distribution and Risk or Mean Differences of Observed Confounders by Oncotype Score Threshold for Patients with Oncotype Scores in Window [25,26]

	Oncotype Score < 26	Oncotype Score ≥ 26	Difference	p-value
Age	58.42	56.26	-2.16	0.49
White	0.92	0.96	0.04	0.65
Af-American	0.05	0.04	-0.02	1.00
Other Race	0.03	0.00	-0.03	1.00
Hispanic	0.00	0.00	0.00	1.00
Lo Grade Tumor	0.75	0.67	-0.08	0.54
Int. Grade Tumor	0.25	0.33	0.08	0.76
High Grade Tumor	0.00	0.00	0.00	1.00
Grade Miss	0.26	0.22	-0.04	0.79
Tumor Size	18.82	13.11	-5.70	0.04
Lymphovascular Invasion	0.11	0.08	-0.03	0.85
Estrogen Receptor	10.00	10.00	0.00	1.00
Proges. Receptor	11.32	12.22	0.91	0.49
Under 50	0.24	0.33	0.10	0.43
Surgery Type 0/1	0.74	0.67	-0.07	0.60
Tumor Size Missing	0.00	0.00	0.00	1.00

Note: The first two columns show means for each group. The third column shows difference in means between both groups. The last column shows the Fisherian p-value assuming a fixed margins randomization mechanism that assigns units with oncotype score in the window [25,26] to be above or below the 26 cutoff. There are 38 observations below the cutoff, and 27 above the cutoff.

We repeat the analysis for the next larger symmetric window, $W = [24, 27]$, the results are in Table 8. For this larger window, the minimum p-value improves: the two smallest p-values are 0.10 (for tumor size) and 0.08 (for lymphovascular invasion). Despite the modest improvement, both covariates are below the recommended 0.15 p-value threshold, and both of them are potentially important confounders.

Table 8: Distribution and Risk or Mean Differences of Observed Confounders by Oncotype Score Threshold for Patients with Oncotype Scores in Window [,27]

	Oncotype Score < 26	Oncotype Score ≥ 26	Difference	p-value
Age	57.67	57.20	-0.47	0.83
White	0.92	0.86	-0.06	0.36
Af-American	0.07	0.08	0.01	1.00
Other Race	0.01	0.06	0.05	0.28
Hispanic	0.00	0.00	0.00	1.00
Lo Grade Tumor	0.65	0.70	0.05	0.68
Int. Grade Tumor	0.35	0.30	-0.05	0.68
High Grade Tumor	0.00	0.00	0.00	1.00
Grade Miss	0.24	0.20	-0.04	0.64
Tumor Size	16.33	13.36	-2.97	0.10
Lymphovascular Invasion	0.20	0.09	-0.11	0.08
Estrogen Receptor	10.00	10.00	0.00	1.00
Proges. Receptor	11.20	12.20	1.00	0.14
Under 50	0.27	0.26	-0.01	1.00
Surgery Type 0/1	0.71	0.62	-0.09	0.34
Tumor Size Missing	0.00	0.00	0.00	1.00

Note: The first two columns show means for each group. The third column shows difference in means between both groups. The last column shows the Fisherian p-value assuming a fixed margins randomization mechanism that assigns units with oncotype score in the window [24,27] to be above or below the 26 cutoff. There are 83 observations below the cutoff, and 50 above the cutoff.

We also analyze the next three larger windows around the cutoff: $[23, 28]$, $[22, 29]$, and $[21, 30]$. To conserve space, we summarize the results in Figure 5, where we report the minimum p-value obtained for each window. As we can see, the minimum p-value is below 0.03 in all of these 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. The ideal scenario would be to choose a window where all confounders were balanced with a minimum p-value of 0.15 or larger; however, this is not possible in this case. We find that one specific covariate, tumor size, is imbalanced (with p-value below the recommended 0.15) in all the windows we tested. Moreover, we found a second covariate, lymphovascular invasion, that is also imbalanced in the second smallest window. Given these covariate differences, should we proceed with the outcome analysis? For

pedagogical and illustrative purposes, we will show the results for the outcome variable in the two smallest windows, $[25, 26]$ and $[24, 27]$, where only tumor size or both tumor size and lymphovascular invasion are imbalanced. However, the results should be interpreted with caution: these covariates are likely to be important determinants of the outcome, and thus the outcome results are likely to be confounded and fail to provide a valid estimate of the true effect of chemotherapy on cancer recurrence.

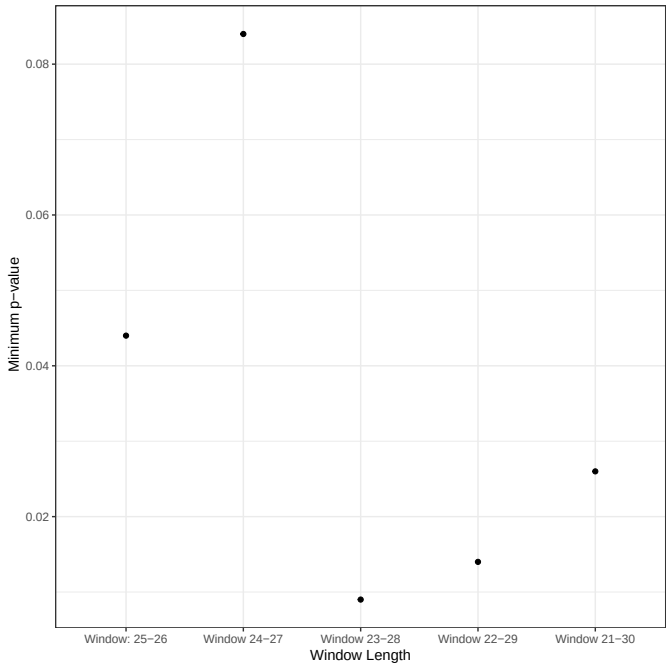


Figure 5: Minium p-value from Fisherian tests of 16 covariates in five symmetric windows.

To start the outcome analysis, we first present the first-stage effect, $\theta_{\mathtt{D}}$, to understand whether reaching an oncotype score of 26 resulted in a significant increase in the probability of being treated with adjuvant chemotherapy. Table 9 contains the results in both windows. We report the first-stage effect, $\theta_{\mathtt{D}}$, and the Fisherian p-value corresponding to the sharp null hypothesis of no effect for any unit. We also report a large-sample analysis based on the F-statistic, which is standard to assess the strength of the instrument. First, although in the smallest window $W = [25, 26]$ the proportion of patients with score equal to 26 who receive chemotherapy is 15 percentage points larger than the proportion of patients with score equal to 25 who receive the treatment, the Fisherian p-value of the hypothesis that the effect is zero for very patient is 0.32. The p-value in the larger window decreases to 0.05, but both F-statistics are small and fail to exceed the usual rule of thumb of 16.

All in all, the evidence shows that the instrument lacks strength: the fact that patients are treated at values smaller than 26 makes the score a weak instrument at the cutoff.

Table 9: First Stage Association Between Oncotype Threshold and Receipt of Chemotherapy

	Window 1 25-26	Window 2 24–27
1st Stage Estimate	0.15	0.19
Fisherian p-value	0.32	0.05
R-squared	0.02	0.03
F-statistic	1.51	4.63
F-test p-value	0.22	0.03
Sample Size	65	133

Note: The row labeled 1st Stage Estimate shows the estimated $\theta_{\mathtt{D}}$, the difference in the percentage of patients receiving chemotherapy when the oncotype score is 26 or above versus when it is below 26, for patients in the respective window. The Fisherian p-value assumes a fixed margins randomization mechanism that assigns units with oncotype score in the respective window to be above or below the 26 cutoff.

We now turn to analyze the intention-to-treat (ITT) effect, $\theta_{\mathtt{Y}}$. It is useful to begin with a graphical analysis, which we previewed in Figure 3. In this plot, we did not observed any obvious relationship between the oncotype score and the outcome. That is, while we observed a modest increase in cancer recurrence for oncotype scores above 26, these estimates appeared to be imprecisely estimated, with very long confidence intervals. We also conduct a formal ITT analysis in both windows; Table 10 contains the results. The magnitude of the estimate is relatively similar in both windows, around 7 percentage points. However, the confidence intervals are long and barely include zero in both cases.

Since this is a fuzzy RD, in addition to the ITT effects, which capture the effects of having an oncotype score of 26 or higher, we are interested in the effects of actually receiving chemotherapy. Given the weakness of the instrument reported in Table 9, however, we must be very cautious about proceeding with the fuzzy RD analysis. The key difference between the fuzzy RD analysis and the ITT analysis is that, subject to the IV assumptions, an ITT analysis estimates the effect of having an oncotype score above 25 on cancer recurrence while the fuzzy RD analysis estimate the effect of

Table 10: Intention-to-Treat RD Estimates of Risk Differences of Recurrence of Breast Cancer for Patients With Oncotype Scores of 25 or Less Compared with Oncotype Scores of 26 or Greater.

	Window	Risk Difference	95% Confidence Interval	p-value
ITT Effect of Oncotype Score above 26 on Breast Cancer	[25, 26]	0.074	[0.00, 0.16]	0.185
ITT Effect of Oncotype Score above 26 on Breast Cancer	[24, 27]	0.068	[0.00, 0.13]	0.053

Note: The rows show the estimated ITT effect, $\widehat{\tau}_{\mathrm{Y}}$, for units with oncotype scores in the windows [25,26] and [24,27]. The window [25,26] has 65 observations, and the window [24,27] has 133 observations. A positive effect indicates higher incidence of recurrence. The last column shows the Fisherian p-value assuming a fixed margins randomization mechanism that assigns units with oncotype score in the respective window to be above or below the 26 cutoff; the confidence interval is obtained by inversion of this Fisherian test.

receiving chemotherapy on cancer recurrence. However, in this setting, it is not advisable to conduct the fuzzy RD analysis given that the instrument is weak in both of the windows considered.

Overall, the conclusions we can reach from this RD design are limited by the fact that the clinical guideline was not followed strongly, which resulted in a very weak instrument, and the fact that important confounders are imbalanced even in the smallest windows around the cutoff suggested in the clinical guideline. In general, the lack of conclusive evidence we find appears consistent with the TAILORx randomized controlled trial (Sparano et al., 2018), which showed no difference in the 5-year or 9-year risk of breast cancer recurrence for patients with oncotype scores within [11,25] who underwent chemotherapy versus endocrine therapy.

5.5.2 Patient Cost-Sharing and Healthcare Utilization

We now present the analysis of the cost-sharing application, our third and final example. We begin with the standard RD plot to illustrate the design, where the data are binned by day. Figure 6 shows a very clear pattern in the data: when children are three or older, the rate of hospital visits drops noticeably.

For the following analyses and for the purposes of illustration, we transform the original running variable (which is days from the third birthday) to be at the weekly rather than daily levels. When the running variable is transformed in this way, every value of the score (which is now weeks from the third birthday) now has seven repeated values coming from the seven days within that week. This pattern of mass points can be seen clearly in Figure 7, where we plot the raw data binned by seven-day periods for a small number of periods around the cutoff. This plot allows us to see that each bin of the score contains multiple observations stacked vertically—this vertical pattern is the

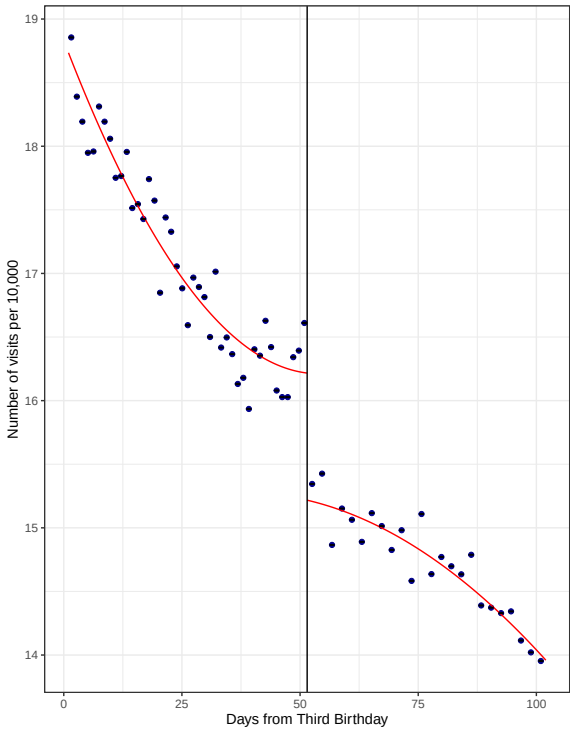


Figure 6: Change in hospital utilization before and after a child's third birthday.

key characteristic of an RD plot when the score has mass points.

Using the weekly score with seven-observation mass points, 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 all the windows contained in it. We use four pre-determined covariates in our window selector: share of male children, household income per capita, share of children born in Taipei, and birth year. Figure 8 shows the results of the window selection, plotting the minimum Fisherian p-value found in ten nested windows, starting with a window one week on either side of the cutoff, and increasing symmetrically by one week every time. 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 11 shows the results of the balance tests in our selected window. In this window, not only are the p-values above a 0.05 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

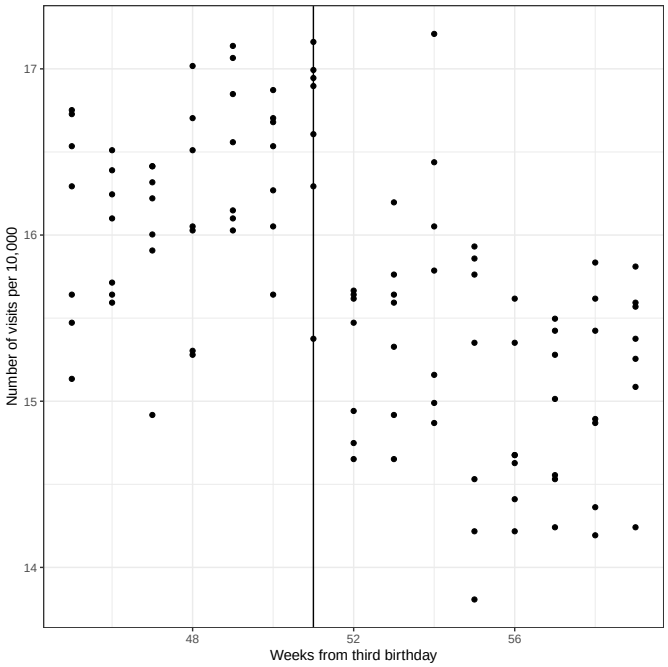


Figure 7: Binning of daily observations be week above and below the cutoff of the third birthday.

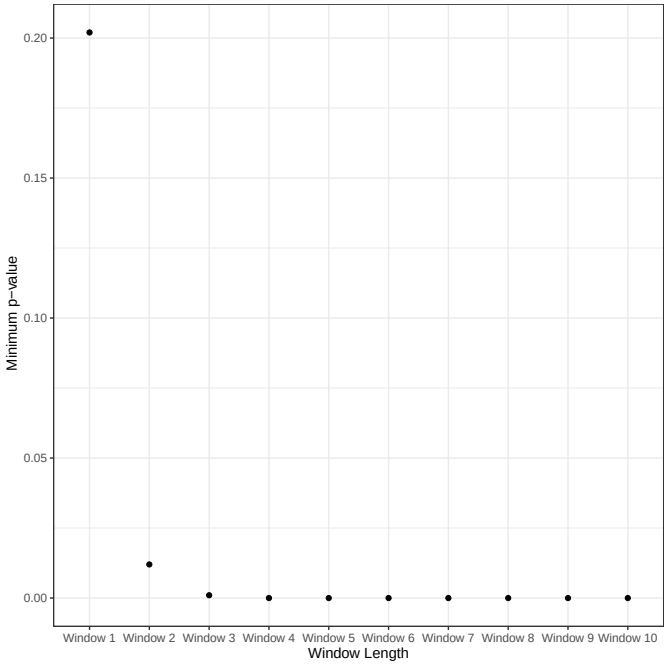


Figure 8: Minium p-value from Fisherian tests of baseline covariates in symmetric windows by weeks.

Table 11: Distribution of Observed Confounders From Data Driven Window Selection: 1 week above and below cutoff

	Mean Below	Mean Above	Diff. in Means	p-value
Share of male	0.55	0.55	0.00	0.78
Household Income per Capita	12494.53	12532.86	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

Notes: 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).

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 approximations and provide exact p-values even when sample sizes are very small as in this case. The results are shown in Table 12, 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%. This suggests, as expected, that cost sharing causes families to use medical care at higher rates (recall that cost-sharing is lost once the child turns three years of age).

Table 12: RD Effect of Cost Sharing on Utilization

	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

Notes: 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).

For robustness, we also ran a continuity-based analysis on the same weekly data (with the running variable aggregated at the weekly level and seven-day mass points), and on the daily data (with the running variable at the daily level and no mass points). With the weekly data, the RD effect is -1.031 with robust confidence interval of $[-1.479, -0.642]$ and robust p-value of zero (main MSE-optimal bandwidth equal to 14.11 weeks); with the daily data, the RD effect is -0.990

with robust confidence interval of $[-1.433, -0.583]$ and robust p-value of zero (main MSE-optimal bandwidth equal to 105.52 days). Thus, the conclusions remain unchanged.

Finally, we conduct a placebo analysis to assess the validity of the design. The replication data contains information on healthcare utilization for the period from $1997 - 2002$. In this time frame, cost-sharing was not higher for children under the age three. Thus, we expect no detectable treatment effects from a similar analysis in this pre-treatment period. Indeed, Table 13 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 post-treatment 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.

Table 13: Placebo RD Effect of Cost Sharing on Utilization: Pre-treatment Period

	Mean Below	Mean Above	Diff. in Means	Fisherian p-value
Number of Doctor Visits per 10,000	11.310	11.183	-0.127	0.209

Notes: 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).

This third application represents the RD design at its ideal. That is, in this application, the design passes all the key falsification tests. As a result, the outcome analysis is straightforward. We also find that the results are robust to any variation in the methods used to estimate the treatment effect. Moreover, these results also underscore the fact that a sharp RD design greatly simplifies the analysis.

6 Conclusion

Our discussion, empirical examples, and accompanying codes, have sought to present a simple and formal guide to practice for those who wish to analyze RD designs in biomedical contexts. Our discussion covered modern validation, estimation, and inference approaches for RD designs based on both continuity and local randomization approaches.

We have discussed new developments in the interpretation and analysis of RD designs, with special emphasis on two complications that are common in biomedical studies: discrete scores

and imperfect compliance. The RD score is discrete when multiple observations share the same value of the score—this is common when clinical guidelines are based on cutoffs of genetic or other coarse scores, as we illustrated with our re-analysis of a study of the effects of chemotherapy on breast cancer recurrence. When the score has such "mass points", the RD analysis has to be modified accordingly. Continuity-based methods are typically not appropriate unless the number of mass points is large or the researcher is willing to invoke parametric models. Local randomization methods are often more appropriate. When the number of observations in each mass point is large, the researcher can simply compare the first treated mass point with the last control mass point, and minimize the need for extrapolation. When the number of observations per mass point is not large enough and several mass points have to be combined for analysis, local randomization methods can investigate the sequence of symmetric windows around the cutoff defined by sequentially adding one extra mass point on either side; this makes the process of window selection simpler than when the score is continuous. When the score is discrete, however, local randomization methods do require modifying the parameter of interest so that the effect compares treated outcomes at the cutoff with control outcomes at the value just before the cutoff; whether this redefinition is sensible should be judged on case-by-case basis.

Imperfect compliance occurs when some units below the cutoff receive the treatment and/or some units above the cutoff fail to receive the treatment. This is a standard feature in biomedical applications, where clinical guidelines are not binding and physicians can decide whether to prescribe a treatment—and where patients can be encouraged but not forced to take medication, as illustrated with our ART application. In this case, several concepts from the analysis of IV become relevant for the RD context, in particular the strength of the instrument. As we illustrated with the oncotype application, when a clinical guideline based on a cutoff is only weakly followed by physicians who recommend a course of treatment, the probability of receiving treatment at the cutoff will not be sufficiently different from the the probability of receiving treatment just below the cutoff. This will cause the instrument to be weak, and will compromise the ability of the RD design to be informative about the effect of the treatment on the outcomes of interest.

Regardless of the approach chosen, we hope our discussion will contribute to a transparent and reproducible analysis of RD designs in biomedical applications.

References

Abadie, A., and Cattaneo, M. D. (2018), "Econometric Methods for Program Evaluation," *Annual Review of Economics*, 10, 465–503.

Albain, K. S., Barlow, W. E., Shak, S., Hortobagyi, G. N., Livingston, R. B., Yeh, I.-T., Ravdin, P., Bugarini, R., Baehner, F. L., Davidson, N. E. et al. (2010), "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," *The Lancet Oncology*, 11, 55–65.

Arai, Y., Hsu, Y., Kitagawa, T., Mourifié, I., and Wan, Y. (2021a), "Testing Identifying Assumptions in Fuzzy Regression Discontinuity Designs," *Quantitative Economics*.

Arai, Y., and Ichimura, H. (2018), "Simultaneous Selection of Optimal Bandwidths for the Sharp Regression Discontinuity Estimator," *Quantitative Economics*, 9, 441–482.

Arai, Y., Otsu, T., and Seo, M. H. (2021b), "Regression Discontinuity Design with Potentially Many Covariates," arXiv:2109.08351.

Baiocchi, M., Cheng, J., and Small, D. S. (2014), "Instrumental variable methods for causal inference," *Statistics in Medicine*, 33, 2297–2340.

Barreca, A. I., Lindo, J. M., and Waddell, G. R. (2016), "Heaping-Induced Bias in Regression-Discontinuity Designs," *Economic Inquiry*, 54, 268–293.

Boon, M. H., Craig, P., Thomson, H., Campbell, M., and Moore, L. (2021), "Regression discontinuity designs in health: a systematic review," *Epidemiology (Cambridge, Mass.)*, 32, 87.

Bor, J., Fox, M. P., Rosen, S., Venkataramani, A., Tanser, F., Pillay, D., and Bärnighausen, T. (2017), "Treatment eligibility and retention in clinical HIV care: A regression discontinuity study in South Africa," *PLoS Medicine*, 14, e1002463.

Bor, J., Moscoe, E., and Bärnighausen, T. (2015), "Three approaches to causal inference in regression discontinuity designs," *Epidemiology*, 26, e28–e30.

Bor, J., Moscoe, E., Mutevedzi, P., Newell, M.-L., and Bärnighausen, T. (2014), "Regression Discontinuity Designs in Epidemiology: Causal Inference without Randomized Trials," *Epidemiology*, 25, 729–737.

Calonico, S., Cattaneo, M. D., and Farrell, M. H. (2018), "On the Effect of Bias Estimation on Coverage Accuracy in Nonparametric Inference," *Journal of the American Statistical Association*, 113, 767–779.

——— (2020), "Optimal Bandwidth Choice for Robust Bias Corrected Inference in Regression Discontinuity Designs," *Econometrics Journal*, 23, 192–210.

——— (2022), "Coverage Error Optimal Confidence Intervals for Local Polynomial Regression," *Bernoulli,* forthcoming.

Calonico, S., Cattaneo, M. D., Farrell, M. H., and Titiunik, R. (2019), "Regression Discontinuity Designs using Covariates," *Review of Economics and Statistics*, 101, 442–451.

Calonico, S., Cattaneo, M. D., and Titiunik, R. (2014), "Robust Nonparametric Confidence Intervals for Regression-Discontinuity Designs," *Econometrica*, 82, 2295–2326.

——— (2015), "Optimal Data-Driven Regression Discontinuity Plots," *Journal of the American Statistical Association*, 110, 1753–1769.

Cattaneo, M. D., Frandsen, B., and Titiunik, R. (2015), "Randomization Inference in the Regression Discontinuity Design: An Application to Party Advantages in the U.S. Senate," *Journal of Causal Inference*, 3, 1–24.

Cattaneo, M. D., Idrobo, N., and Titiunik, R. (2020a), *A Practical Introduction to Regression Discontinuity Designs: Foundations*, Cambridge Elements: Quantitative and Computational Methods for Social Science, Cambridge University Press.

——— (2022), *A Practical Introduction to Regression Discontinuity Designs: Extensions*, Cambridge Elements: Quantitative and Computational Methods for Social Science, Cambridge University Press*, to appear.*

Cattaneo, M. D., Jansson, M., and Ma, X. (2020b), "Simple Local Polynomial Density Estimators," *Journal of the American Statistical Association*, 115, 1449–1455.

Cattaneo, M. D., Keele, L., Titiunik, R., and Vazquez-Bare, G. (2016), "Interpreting Regression Discontinuity Designs with Multiple Cutoffs," *Journal of Politics*, 78, 1229–1248.

——— (2021), "Extrapolating Treatment Effects in Multi-Cutoff Regression Discontinuity Designs," *Journal of the American Statistical Association*, 116, 1941–1952.

Cattaneo, M. D., and Titiunik, R. (2022), "Regression Discontinuity Designs," *Annual Review of Economics.*

Cattaneo, M. D., Titiunik, R., and Vazquez-Bare, G. (2017), "Comparing Inference Approaches for RD Designs: A Reexamination of the Effect of Head Start on Child Mortality," *Journal of Policy Analysis and Management*, 36, 643–681.

Cattaneo, M. D., Titiunik, R., and Vazquez-Bare, G. (2020c), "The Regression Discontinuity Design," in *Handbook of Research Methods in Political Science and International Relations*, eds. L. Curini and R. J. Franzese, Sage Publications, pp. 835–857.

Cattaneo, M. D., and Vazquez-Bare, G. (2016), "The Choice of Neighborhood in Regression Discontinuity Designs," *Observational Studies*, 2, 134–146.

Craig, P., Katikireddi, S. V., Leyland, A., and Popham, F. (2017), "Natural Experiments: An Overview of Methods, Approaches, and Contributions to Public Health Intervention Research," *Annual Review of Public Health*, 38, 39–56.

Davies, N. M., Thomas, K. H., Taylor, A. E., Taylor, G. M., Martin, R. M., Munafò, M. R., and Windmeijer, F. (2017), "How to compare instrumental variable and conventional regression analyses using negative controls and bias plots," *International Journal of Epidemiology*, 46, 2067–2077.

De Magalhães, L., Hangartner, D., Hirvonen, S., Meriläinen, J., Ruiz, N., and Tukiainen, J. (2020), "How Much Should We Trust Regression Discontinuity Design Estimates? Evidence from Experimental Benchmarks of the Incumbency Advantage," *working paper*.

Dong, Y. (2015), "Regression Discontinuity Applications with Rounding Errors in the Running Variable," *Journal of Applied Econometrics*, 30, 422–446.

——— (2018), "Alternative Assumptions to Identify LATE in Fuzzy Regression Discontinuity Designs," *Oxford Bulletin of Economics and Statistics*, 80, 1020–1027.

Dong, Y., Lee, Y.-Y., and Gou, M. (2021), "Regression Discontinuity Designs with a Continuous Treatment," *Journal of the American Statistical Association*, forthcoming.

Ernst, M. D. (2004), "Permutation Methods: A Basis for Exact Inference," *Statistical Science*, 19, 676–685.

Fan, J., and Gijbels, I. (1996), *Local polynomial Modelling and Its Applications*, Vol. 66, CRC Press.

Feir, D., Lemieux, T., and Marmer, V. (2016), "Weak identification in fuzzy regression discontinuity designs," *Journal of Business & Economic Statistics*, 34, 185–196.

Ganong, P., and Jäger, S. (2018), "A Permutation Test for the Regression Kink Design," *Journal of the American Statistical Association*, 113, 494–504.

Garabedian, L. F., Chu, P., Toh, S., Zaslavsky, A. M., and Soumerai, S. B. (2014), "Potential bias of instrumental variable analyses for observational comparative effectiveness research," *Annals of Internal Medicine*, 161, 131–138.

Gelman, A., and Imbens, G. W. (2019), "Why High-Order Polynomials Should Not be Used in Regression Discontinuity Designs," *Journal of Business & Economic Statistics*, 37, 447–456.

Glymour, M. M., Tchetgen Tchetgen, E. J., and Robins, J. M. (2012), "Credible Mendelian randomization studies: Approaches for evaluating the instrumental variable assumptions," *American Journal of Epidemiology*, 175, 332–339.

Hahn, J., Todd, P., and van der Klaauw, W. (2001), "Identification and Estimation of Treatment Effects with a Regression-Discontinuity Design," *Econometrica*, 69, 201–209.

Han, H.-W., Lien, H.-M., and Yang, T.-T. (2020), "Patient Cost-Sharing and Healthcare Utilization in Early Childhood: Evidence from a Regression Discontinuity Design," *American Economic Journal: Economic Policy*, 12, 238–78.

Hernán, M. A. (2018), "The C-word: Scientific Euphemisms Do Not Improve Causal Inference from Observational Data," *American Journal of Public Health*, 108, 616–619.

Hernán, M. A., and Robins, J. M. (2022), *Causal Inference: What If*, Boca Raton: Chapman & Hall/CRC.

Houlihan, C. F., Bland, R. M., Mutevedzi, P. C., Lessells, R. J., Ndirangu, J., Thulare, H., and Newell, M.-L. (2010), "Cohort Profile: Hlabisa HIV Treatment and Care Programme," *International Journal of Epidemiology*, 40, 318–326.

Hyytinen, A., Meriläinen, J., Saarimaa, T., Toivanen, O., and Tukiainen, J. (2018), "When Does Regression Discontinuity Design Work? Evidence from Random Election Outcomes," *Quantitative Economics*, 9, 1019–1051.

Imbens, G. W., and Rosenbaum, P. (2005), "Robust, Accurate Confidence Intervals with a Weak Instrument: Quarter of Birth and Education," *Journal of The Royal Statistical Society Series A*, 168, 109–126.

Imbens, G. W., and Rubin, D. B. (2015), *Causal Inference in Statistics, Social, and Biomedical Sciences*, Cambridge University Press.

Kamat, V. (2018), "On Nonparametric Inference in the Regression Discontinuity Design," *Econometric Theory*, 34, 694–703.

Kang, H., Peck, L., and Keele, L. (2018), "Inference for Instrumental Variables: A Randomization Inference Approach," *Journal of The Royal Statistical Society, Series A*, 181, 1231–1254.

Kaptchuk, T. J., and Miller, F. G. (2015), "Placebo effects in medicine," *N Engl J Med*, 373, 8–9.

Keele, L. J., Small, D. S., and Grieve, R. (2017), "Randomization Based Instrumental Variables Methods for Binary Outcomes with an Application to the IMPROVE Trial," *Journal of The Royal Statistical Society, Series A*, 180, 569–586.

Keele, L. J., and Titiunik, R. (2015), "Geographic Boundaries as Regression Discontinuities," *Political Analysis*, 23, 127–155.

Keele, L. J., Titiunik, R., and Zubizarreta, J. (2015), "Enhancing a Geographic Regression Discontinuity Design Through Matching to Estimate the Effect of Ballot Initiatives on Voter Turnout," *Journal of the Royal Statistical Society: Series A*, 178, 223–239.

Keele, L. J., Zhao, Q., Kelz, R. R., and Small, D. S. (2019), "Falsification Tests for Instrumental Variable Designs with an Application to Tendency to Operate," *Medical Care*, 57, 167–171.

Korting, C., Lieberman, C., Matsudaira, J., Pei, Z., and Shen, Y. (2022), "Visual Inference and Graphical Representation in Regression Discontinuity Designs," *arXiv preprint arXiv:2112.03096*.

Lee, D. S. (2008), "Randomized Experiments from Non-random Selection in U.S. House Elections," *Journal of Econometrics*, 142, 675–697.

Li, F., Mattei, A., and Mealli, F. (2015), "Evaluating the Causal Effect of University Grants on Student Dropout: Evidence from a Regression Discontinuity Design using Principal Stratification," *Annals of Applied Statistics*, 9, 1906–1931.

Maciejewski, M. L., and Basu, A. (2020), "Regression Discontinuity Design," *JAMA*, 324, 381–382.

McCrary, J. (2008), "Manipulation of the running variable in the regression discontinuity design: A density test," *Journal of Econometrics*, 142, 698–714.

O'Keeffe, A. G., Geneletti, S., Baio, G., Sharples, L. D., Nazareth, I., and Petersen, I. (2014), "Regression Discontinuity Designs: An Approach to the Evaluation of Treatment Efficacy in Primary Care Using Observational Data," *BMJ*, 349:g5293.

Paik, S., Tang, G., Shak, S., Kim, C., Baker, J., Kim, W., Cronin, M., Baehner, F. L., Watson, D., Bryant, J., Costantino, J. P., Geyer, C. E. J., Wickerham, D. L., and Wolmark, N. (2006), "Gene Expression and Benefit of Chemotherapy in Women With Node-Negative, Estrogen Receptor–Positive Breast Cancer," *Journal of Clinical Oncology*, 24, 3726–3734.

Papay, J. P., Willett, J. B., and Murnane, R. J. (2011), "Extending the regression-discontinuity approach to multiple assignment variables," *Journal of Econometrics*, 161, 203–207.

Pei, Z., Lee, D. S., Card, D., and Weber, A. (2021), "Local Polynomial Order in Regression Discontinuity Designs," *Journal of Business & Economic Statistics*, 31, 507–524.

Pizer, S. D. (2016), "Falsification testing of instrumental variables methods for comparative effectiveness research," *Health Services Research*, 51, 790–811.

Reardon, S. F., and Robinson, J. P. (2012), "Regression discontinuity designs with multiple rating-score variables," *Journal of Research on Educational Effectiveness*, 5, 83–104.

Rosenbaum, P. R. (2010), *Design of Observational Studies*, Springer.

Sekhon, J. S., and Titiunik, R. (2016), "Understanding Regression Discontinuity Designs as Observational Studies," *Observational Studies*, 2, 174–182.

——— (2017), "On Interpreting the Regression Discontinuity Design as a Local Experiment," in *Regression Discontinuity Designs: Theory and Applications (Advances in Econometrics, volume 38)*, eds. M. D. Cattaneo and J. C. Escanciano, Emerald Group Publishing, pp. 1–28.

Sparano, J. A., Gray, R. J. ., Makower, D. F., Pritchard, K. I., Albain, K. S., Hayes, D. F., Jr, C. E. G., Dees, E. C., Perez, E. A., Jr, J. A. O. et al. (2018), "Adjuvant chemotherapy guided by a 21-gene expression assay in breast cancer," *New England Journal of Medicine*, 379.

Sparano, J. A., and Paik, S. (2008), "Development of the 21-Gene Assay and Its Application in Clinical Practice and Clinical Trials," *Journal of Clinical Oncology*, 26, 721–728.

Swanson, S. A., and Hernán, M. A. (2013), "Commentary: How to Report Instrumental Variable Analyses (Suggestions Welcome)," *Epidemiology*, 24, 370–374.

Tanser, F., Hosegood, V., Bärnighausen, T., Herbst, K., Nyirenda, M., Muhwava, W., Newell, C., Viljoen, J., Mutevedzi, T., and Newell, M.-L. (2007), "Cohort Profile: Africa Centre Demographic Information System (ACDIS) and Population-based HIV Survey," *International Journal of Epidemiology*, 37, 956–962.

Thistlethwaite, D. L., and Campbell, D. T. (1960), "Regression-Discontinuity Analysis: An Alternative to the Ex-Post Facto Experiment," *Journal of Educational Psychology*, 51, 309–317.

Titiunik, R. (2021), "Natural Experiments," in *Advances in Experimental Political Science*, eds. J. N. Druckman and D. P. Gree, chapter 6, Cambridge University Press, pp. 103–129.

Tuvaandorj, P. (2020), "Regression Discontinuity Designs, White Noise Models, and Minimax," *Journal of Econometrics*, 218, 587–608.

Xu, K.-L. (2017), "Regression Discontinuity with Categorical Outcomes," *Journal of Econometrics*, 201, 1–18.

Zhao, Q., and Small, D. S. (2018), "Graphical diagnosis of confounding bias in instrumental variables analysis," *Epidemiology*, 29, e29–e31.