Regression Discontinuity Designs in the Evaluation of Section 1115 Demonstrations

White Paper

February 2023

Monica Farid, Wenjia Zhu, and Anna Hill

This white paper was prepared on behalf of the Centers for Medicare & Medicaid Services (CMS) as part of the Medicaid 1115 Demonstration Support Contract (contract number: HHSM-500-2014-00034I/75FCMC19F0008). Under the contract, Mathematica provides technical assistance focused on states' section 1115 demonstration evaluation designs and reports. This paper is intended to support states and their evaluators by providing a nontechnical summary of best practices for supporting rigorous implementations of regression discontinuity design.

Contents

I. Introduction 1

II. Overview of RDD 3

A. RDD assumptions 3

B. Interpretation of RDD estimates 4

III. Best Practices 6

A. Best practices for graphical representation of a discontinuity 6

B. Choosing the functional form of the regression 8

C. Choosing the bandwidth around the eligibility threshold and the weighting scheme 9

D. Choosing the functional form and bandwidth simultaneously 10

E. Confidence intervals for valid inference 10

IV. Sensitivity Analyses 11

V. Robustness Checks 13

A. Checking for balance on covariates and placebo outcomes near the threshold 13

B. Investigating the density of the running variable 13

C. Using artificial thresholds (or placebo thresholds) 14

VI. Extensions 17

A. Fuzzy RDD 17

B. Local randomization approach to RDD 18

VII. Conclusion 21

References 22

I. Introduction

Medicaid section 1115 demonstrations often use thresholds or cut-offs based on continuous variables such as age or income to determine whether beneficiaries are subject to certain policies.[1] Several section 1115 demonstration components apply to those older than age 20 or apply only to those with incomes between 100 and 138 percent of the federal poverty level (FPL). For example, New Jersey's substance use disorder (SUD) demonstration, approved within the state's broader section 1115 demonstration, "FamilyCare Comprehensive Demonstration", for the performance period August 1, 2017 through June 30, 2022 expanded SUD services in institutions of mental disease (IMDs) to beneficiaries aged 21 and above.[2]

Eligibility thresholds present an opportunity for states to evaluate the causal impact of demonstration components using a regression discontinuity design (RDD). RDD takes advantage of the likely similarity of people who are within a small window or bandwidth around the eligibility threshold. When eligibility for a section 1115 demonstration policy is based on an income threshold of 133 percent of the FPL, an RDD might consider people with incomes between 128 and 138 percent of the FPL.[3] RDD-based analyses compare average outcomes of beneficiaries who barely meet the eligibility criteria to the outcomes of beneficiaries who just miss the eligibility threshold. The RDD approach produces causal estimates when the beneficiaries are similar on either side of the threshold in every aspect except eligibility for the demonstration. Several states proposed RDD as an analytic approach to assess the causal impact of demonstration components for which eligibility is based on an age, income or a medical risk score threshold. For example, Arkansas implemented an RDD using a risk score threshold to assess the impact of Qualified Health Plans (QHPs) on access to care and health care outcomes.[4,5] Similarly, New Jersey proposed implementing an RDD using an age-based eligibility threshold to investigate the impact of SUD service provision in IMDs.[6]

RDD has several advantages in terms of data requirements. It can provide causal impact estimates even if there is no available comparison group unaffected by the demonstration (either from a different eligibility group within the state or an out-of-state group).[7] It also does not require pre-implementation time series data (as is required in a difference-in-differences or synthetic control design).

---

[1] The methods described in this white paper apply regardless of the unit of analysis (Medicaid beneficiary, health care provider, or managed care entity), but for readability, we refer to beneficiaries when describing regression discontinuity design (RDD) methods generically.

[2] Approved evaluation designs and summative evaluation reports are posted to the administrative record for each section 1115 demonstration: https://www.medicaid.gov/medicaid/section-1115-demo/demonstration-and-waiver-list/index.html

[3] This general principle applies to all eligibility thresholds including those with income disregards. For example, when eligibility is based on an income threshold of 133 percent of the FPL with a 5 percent income disregard, RDD would consider individuals just above and just below the effective threshold of 138 percent of the FPL. In this case, an RDD might consider people with incomes between 133 and 143 percent of the FPL.

[4] Arkansas Health Care Independence Program (demonstration period October 1, 2013 through December 31, 2016; approved evaluation design dated March 24, 2014; summative evaluation report dated June 30, 2018).

[5] Individuals with a composite score of less than 0.18 on a health care needs assessment were assigned to a QHP, while those with a score of 0.18 or higher were assigned to a Medicaid plan.

[6] New Jersey FamilyCare Comprehensive Demonstration, SUD Demonstration (demonstration period August 1, 2017 through June 30, 2022; approved evaluation design dated January 3, 2020).

[7] See "Selecting the Best Comparison Group and Evaluation Design: A Guidance Document for State Section 1115

Estimating accurate causal impacts using RDD requires careful and rigorous implementation of RDD. Evaluators must choose a regression function to fit the data, a bandwidth in which the model will be estimated, and a weighting scheme for observations within the bandwidth. In addition, evaluators must test the plausibility of the assumptions that underlie RDD and conduct sensitivity analyses to assess the robustness of estimates to reasonable alternative design choices.

A recent and fast-expanding literature implements RDD in a range of policy settings and develops best practices in RDD implementation and inference. This literature includes several studies assessing aspects of Medicaid policy. For example, Card and Shore-Sheppard (2004) examined the impact of federal legislative changes in Medicaid eligibility rules in the 1980s on insurance coverage of children from families with low incomes. The authors compared children on either side of the income eligibility threshold to estimate the impact of federal Medicaid expansions. Exploiting similar discrete changes in required premiums by family income, Finkelstein et al. (2019) estimated willingness to pay for health insurance and the cost of health insurance in Massachusetts's Commonwealth Care for adults with low incomes.

This white paper provides states with an accessible resource to support sound implementation of RDD in their evaluations of section 1115 demonstrations.[8] In the following sections, we (1) give an overview of RDD using section 1115 demonstration policies as examples; (2) provide a nontechnical summary of best practices; (3) describe ways to check model assumptions and sensitivity to key design choices; and (4) describe recent extensions to the method that might be useful for evaluating section 1115 demonstrations.

Medicaid Section 1115 Demonstrations

Medicaid is a health insurance program that serves low-income children, adults, individuals with disabilities, and seniors. Medicaid is administered by states and is jointly funded by states and the federal government. Within a framework established by federal statutes, regulations and guidance, states can choose how to design aspects of their Medicaid programs, such as benefit packages and provider reimbursement. Although federal guidelines may impose some uniformity across states, federal law also specifically authorizes experimentation by state Medicaid programs through section 1115 of the Social Security Act. Under section 1115 provisions, states may apply for federal permission to implement and test new approaches to administering Medicaid programs that depart from existing federal rules yet are consistent with the overall goals of the program, likely to meet the objectives of Medicaid, and budget neutral to the federal government.

Demonstration Evaluations" (Bradley et al. 2020) and "Selection of Out-of-State Comparison Groups and the Synthetic Control Method" (Pohl and Bradley 2020) at https://www.medicaid.gov/medicaid/section-1115-demonstrations/1115-demonstration-monitoring-evaluation/1115-demonstration-state-monitoring-evaluation-resources/index.html.

[8] For general guidance on conducting causal evaluations of Section 1115 demonstrations, see "Best Practices in Causal Inference for Evaluations of Section 1115 Eligibility and Coverage Demonstrations" (Contreary et al. 2018) at https://www.medicaid.gov/medicaid/section-1115-demonstrations/1115-demonstration-monitoring-evaluation/1115-demonstration-state-monitoring-evaluation-resources/index.html.

II. Overview of RDD

RDD can be a powerful empirical approach for evaluations of section 1115 demonstrations that determine eligibility using thresholds based on continuous variables such as age or income. Although policymakers do not arbitrarily choose eligibility thresholds for section 1115 demonstrations, these thresholds could be considered as good as arbitrary within the group of people who are close to the threshold. For example, adults who are only a few months below or above an age-based threshold for demonstration services and thus deemed ineligible for the services are likely similar to adults who are eligible by falling within a narrow margin of the same eligibility threshold. It is reasonable to expect that both groups would experience similar outcomes if they both participated in the demonstration or if they both did not participate. Therefore, the group of adults who just miss an eligibility threshold who did not participate in the demonstration might be a good comparison for treated adults who just meet the threshold.

An RDD involves choosing a margin around the eligibility threshold, called the bandwidth, and comparing outcomes between people on either side of the threshold. RDD estimates are typically calculated using regression models, which Section III.B describes. In an RDD, the continuous measure used to define eligibility is usually called the running variable but is sometimes also called a scoring variable. There are several approaches in the literature to conducting RDD and this guide provides a framework for conducting a sharp RDD analysis. Sharp RDD is appropriate in settings where the eligibility threshold or cutoff perfectly predicts whether people are affected by the demonstration. This is likely to be true for most section 1115 demonstration evaluations with eligibility thresholds. In the sharp RDD setting, eligibility and treatment are the same. However, there are settings where some eligible beneficiaries do not enroll in a demonstration or receive demonstration services even if they meet the eligibility threshold. For example, beneficiaries above a health risk score threshold may be eligible to receive certain demonstration services like additional chronic care management supports. However, because of a lack of awareness or information, not all beneficiaries may participate. Several RDD variants account for contexts in which the eligibility threshold imperfectly predicts treatment status, including fuzzy RDD. Fuzzy RDD may be especially helpful in evaluating section 1115 demonstrations affected by the COVID-19 Public Health Emergency (PHE) where the barriers to receiving in-person care and beneficiary hesitation to engage with the health care system may have reduced take-up of services among eligible populations. Section VI provides a brief description of fuzzy RDD and other extensions.

Recommended technical RDD resources

This white paper provides a nontechnical overview of the implementation of an RDD for evaluating section 1115 demonstration, several high-quality technical guides are available for evaluators who intend to implement an RDD and would like more detailed technical information. Evaluators can consult the following for additional detail: “A Practical Introduction to Regression Discontinuity Designs: Foundations” (Cattaneo et al. 2020b), “A Practical Introduction to Regression Discontinuity Designs: Extensions” (Cattaneo and Titiunik 2022),“A Practical Guide to Regression Discontinuity” (Jacob et al. 2012), and “Regression Discontinuity Designs in Economics” (Lee and Lemieux 2010).

A. RDD assumptions

The following assumptions must hold for RDD estimates to be causal—that is, to accurately capture the true effect of treatment (impact of demonstration participation) on the outcomes of interest:

1. **The running variable (the continuous measure used to define eligibility) cannot be manipulated.** For example, in some section 1115 demonstrations, an income threshold determines whether a beneficiary qualifies for premium assistance. The RDD assumption is violated if beneficiaries can strategically lower their income to receive premium assistance. Those close to but on either side of the threshold might no longer be comparable to one another because beneficiaries who strategically lower their income might have different characteristics than those who do not. Age is a running variable that is impossible to manipulate, so this assumption would almost certainly hold in RDD evaluations of section 1115 demonstrations with age-based eligibility thresholds.
2. **All characteristics of people that affect the outcome measure are continuous across the eligibility threshold.** There should not be any systematic differences between people just meeting and just missing the eligibility threshold—that is, they should only differ in their eligibility status. This assumption could be violated if other important policy changes occur at the eligibility threshold. For example, an evaluator examining changes in outcomes just above and just below an income eligibility threshold that equals 100 percent FPL should be aware that other important public assistance programs might use the same income threshold to determine eligibility. If these other programs affect the outcome of interest, then RDD estimates might be biased.
3. **The outcome variable is not discontinuous at other values of the running variable within the bandwidth.** If discontinuities in the outcome variable are present at other points of the running variable within the bandwidth, it is more difficult to attribute the discontinuity in the outcome at the eligibility threshold to the treatment. For example, a section 1115 demonstration might have an age-based threshold of 20 years for substance use recovery services, and evaluators might want to understand the effect of these services on drug misuse rates. If the drug misuse rate jumps at age 18 when tobacco can be purchased legally or at age 21 when alcohol can be purchased legally, it might be difficult to attribute a change in misuse rates at the eligibility threshold of age 20 to the recovery services.[9]

Although it is difficult to prove that these assumptions hold, evaluators can and should check for violations of these assumptions when conducting analyses with RDD. Section V describes robustness checks that test for violations of RDD assumptions.

Finally, traditional RDD requires that the running variable be continuous; an RDD is not appropriate in settings where eligibility is based on a binary or categorical state like sex or diagnosis. For example, a section 1115 serious mental illness (SMI) demonstration that defines eligibility based on specific diagnoses for SMI could not be evaluated using an RDD.

B. Interpretation of RDD estimates

Although carefully implementing an RDD enables evaluators to estimate the causal impact of section 1115 demonstration services, RDD estimates might not generalize to the entire policy-relevant population (Lee and Lemieux 2010). On the one hand, because RDD relies on the similarity of people close to the eligibility threshold, causal estimates are specific to this small group of people who might not be

[9] However, if the discontinuity that does not occur at the threshold is expected or can be explained by the evaluators, RDD can continue to be appropriate, provided that the outcome variable is continuous for a bandwidth around the threshold that is sufficiently large to allow for precise estimation. To carry on with the example, evaluators could consider a bandwidth within the range of ages 19.5 to 20.5 years.

representative of everyone eligible for the policy. On the other hand, RDD estimates might be especially useful and informative to policymakers considering eligibility changes near the threshold.

Evaluators should be aware of these considerations if they are deciding between RDD and other empirical strategies. If evaluators decide to use an RDD, they should include a discussion of generalizability when describing and discussing the results of an RDD analysis. In some cases, evaluators might be able to use both RDD and other research designs, which would enable them to triangulate estimates of demonstration impacts.

III. Best Practices

Successfully implementing an RDD involves several key design choices. The following sections describe best practices for graphical representation in the RDD setting, choosing the functional form for the regression model, choosing a bandwidth around the discontinuity for the analysis, and estimating confidence intervals that ensure valid inference. Best practices for RDD have evolved considerably over the past two decades, and the recommendations described here might differ from those in older resources and research on this topic.

A. Best practices for graphical representation of a discontinuity

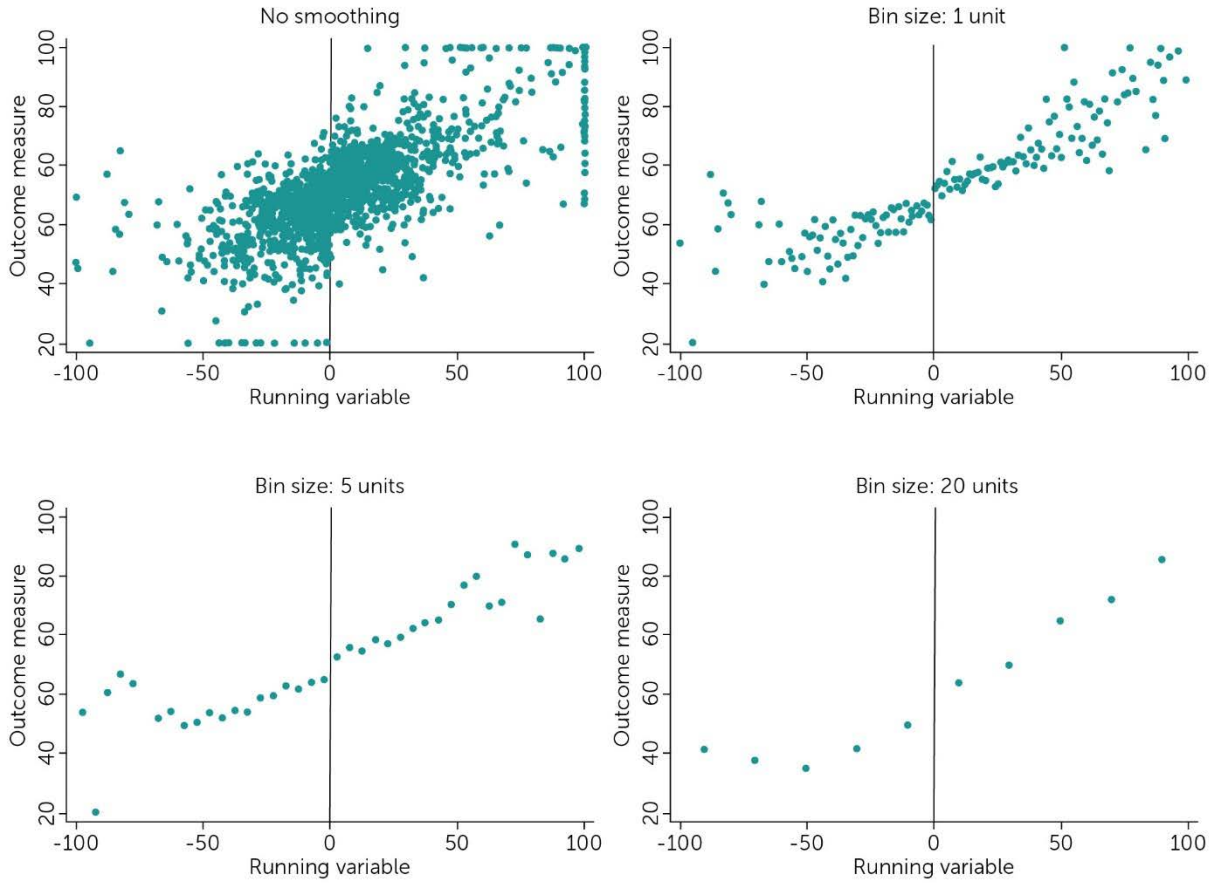
Goals of graphical representation. Graphical representation is a key aspect of RDD. Figures visually convey the size of the discontinuity to readers and provide useful information for evaluators as they make decisions about the empirical approach. Evaluators can use graphs to understand the relationships (1) between the running variable and the likelihood of receiving treatment and (2) between the running variable and the outcome. In particular, RDD plots help evaluators determine the following:

1. Whether a discontinuity exists in the outcome measure at the threshold (see Figure III.1 for plots of an outcome variable against a running variable that depict a discontinuity). Evaluators should graph the outcome measure against the running variable, looking for a visible discontinuity in the outcome at the threshold. If a discontinuity is not visible, it is unlikely that evaluators will find a detectable impact of the demonstration on the outcome measure using RDD.
2. Whether the analysis should be conducted using a sharp or a fuzzy RDD. Evaluators should graph the fraction of eligible beneficiaries participating or engaging in the demonstration component against the running variable and look for a visible discontinuity in this fraction at the eligibility threshold. If not all beneficiaries who meet the threshold are participating in the demonstration, a fuzzy RDD approach might be more appropriate (see Section VI.A for further detail on fuzzy RDD).
3. Reasonable functional forms to describe the relationship between the outcome or treatment variable and running variables. Visual inspection of trends can help evaluators make appropriate choices about the functional forms used to model the relationship between the outcome or treatment measure and the running variable. Section III.B provides more detail on choosing the regression functional form.

Smoothing data and choosing bin sizes. Plotting the outcome or treatment variable against the running variable often involves smoothing. One smoothing method is to group the running variable into bins and plot the average values of the other variables within each bin. If evaluators do not smooth, the scatter plot of the outcome and running variable might be noisy, making it difficult to observe any discontinuity at the threshold (Figure III.1). Smoothing enables evaluators to visually inspect relationships between variables. However, too much smoothing (that is, using bin sizes that are too large) can obscure the true pattern in the data and could even make a discontinuity difficult to observe. The choice of bin size must therefore balance these two considerations.

Figure III.1 depicts an outcome variable plotted against a running variable with no smoothing and with smoothing using three different bin sizes. The top left panel shows no smoothing of the data, and it is difficult to observe any discontinuity. In contrast, the graph in the bottom right panel smooths the data too much, and the outcome variable appears to have no discontinuity in trend across the threshold. The top right and the bottom left panels reveal that a discontinuity does exist at the threshold.

Figure III.1. Smoothing data



Source: Authors' calculations using example data available from the `rdrobust` package in Stata.

Notes: This figure shows an outcome variable plotted against a running variable in four ways. The figure shows a plot with no smoothing and no discontinuity visible (top left), a plot with the outcome variable grouped by 1-unit intervals of the running variable (top right), a plot with the outcome grouped by 5-unit intervals of the running variable (bottom left), and a plot with the outcome grouped by 20-unit intervals of the running variable (bottom right). Moving from a bin size of 1 to 5 improves the usefulness of the graph: the discontinuity where the running variable equals zero is apparent, and the graph shows enough detail to discern the rough relationship between the two variables. The last panel shows that a bin size of 20 units over-smooths the data, obscuring the discontinuity and the shape of the relationship between the two variables.

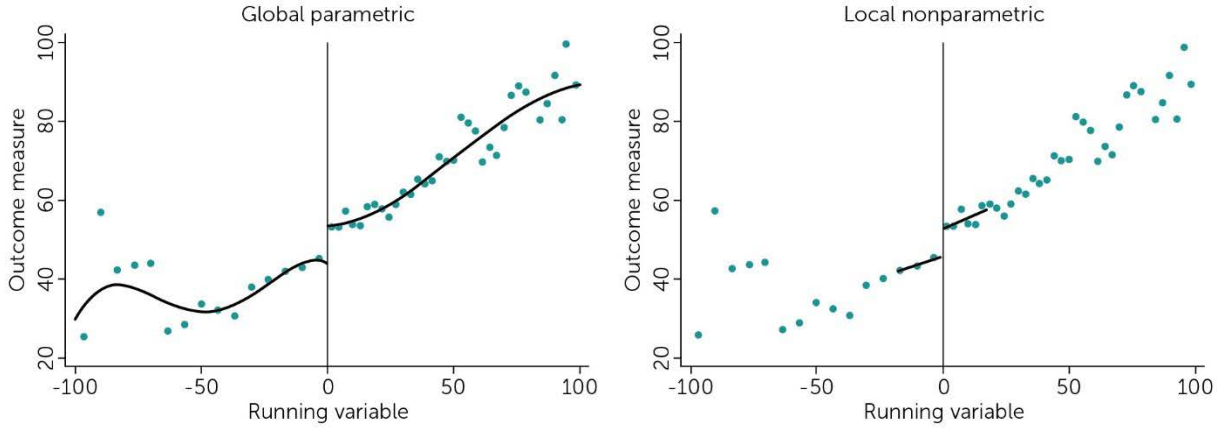
To determine the optimal bin size, evaluators should use an informal and a formal method. The informal method involves choosing a variety of bin sizes, plotting the outcome or treatment variable against the running variable, and inspecting the figures. Through visual inspection, evaluators can usually choose a bin size that is small enough to enable them to observe patterns in the data but wide enough to minimize capturing random noise (see Figure III.1). Using Figure III.1 as an example, an evaluator might choose to test a range of bin sizes close to 5.

One formal method for selecting bin size checks whether a given bin size is too large by comparing the fit of two models (Jacob et al. 2012).[10] In the first model, the outcome variable is regressed on a set of indicators for bins of the running variable using the selected bin size. In the second model, the outcome variable is regressed on indicators for bins of the running variable using half the selected bin size (twice as many indicator variables). If the second model has better fit, it is likely that the selected bin size is too large and smooths the data too much, losing important information. Informed by the visual inspection, evaluators could repeat this process for a range of bin sizes and use the largest of the bin sizes that does not smooth the data too much in the test to produce planned figures.[11]

B. Choosing the functional form of the regression

Approaches to estimating an RDD regression fall into two general categories: global parametric regression and local nonparametric regression. The first approach involves fitting a model to all the observations in the data to estimate the size of the discontinuity in the outcome variable (left panel of Figure III.2), whereas the second only uses data within a small bandwidth around the threshold (right panel of Figure III.2).

Figure III.2. Global parametric and local nonparametric models



Source: Authors' calculations using example data available from the `rdrobust` package in Stata.

Notes: The left-hand panel of this figure shows a fourth order polynomial model fit to the example data. In this scenario, the entire sample is used to estimate the jump in the outcome where the running variable equals zero. The right-hand panel of this figure illustrates a local linear model fit to a subset of the example data within a bandwidth (in this case, the bandwidth was 17.7 units).

The advantage of a global parametric regression is increased statistical power from using all available observations rather than a subset of observations close to the eligibility threshold. There are several formal procedures to choose the best functional form for the global regression—that is, the model that

[10] Model fit is a measure of how well a model explains variation in the data. In this context, evaluators usually measure fit using R-squared.

[11] There are several other formal data-driven tests for determining the optimal bin size. The formal test described is straightforward to implement and produces results comparable to those from more computationally intensive methods. Evaluators interested in exploring more tests for optimal bin size can find descriptions and examples in: Calonico et al. (2014); Lee and Lemieux (2010); and Jacob et al. (2012). The `rdrobust` package in Stata also offers several data-driven approaches to choosing bin size.

best explains the data.[12] However, a key disadvantage of a parametric approach is the potential for bias because of specification error. That is, even when evaluators follow best practices in choosing the functional form, global parametric approaches might not fit data near the threshold well and might therefore incorrectly estimate the size of the discontinuity. Local nonparametric approaches are likely to fit data near the threshold better; however, they have lower statistical power because they rely on a smaller set of observations.

Local nonparametric estimation for RDD evaluations is generally recommended because it offers lower potential for biased estimates relative to global parametric approaches (Gelman and Imbens 2019). One of the most popular local nonparametric approaches is local linear regression, which estimates a linear regression on observations within a small interval around the eligibility threshold. Although a strictly nonparametric approach would simply involve comparing means of the outcome on either side of the threshold, a local linear regression is generally preferred because it is a continuity-based approach with a lower potential for bias in most settings (Cattaneo et al. 2020b). The local linear approach does not require that the true relationship between the running variable and the outcome be linear; rather, it relies on the fact that all functions are approximately linear in small enough sections regardless of their overall shape. For example, although the relationship between the outcome and the running variable is not linear in Figure III.2, a linear model reasonably approximates the true relationship between the two variables in a small bandwidth around the threshold. See Section VI.B for a description of situations in which the simple difference in means, called the local randomization approach, is appropriate.

Although local nonparametric regressions are preferred over global parametric regressions, the latter can provide a useful check of the robustness of local regression estimates. See Section IV for more information about sensitivity testing.

C. Choosing the bandwidth around the eligibility threshold and the weighting scheme

When using a local nonparametric approach to RDD such as a local linear regression, evaluators must choose the bandwidth or the interval around the threshold to estimate the model. This choice involves balancing two considerations. Evaluators have to choose a bandwidth that is narrow enough to ensure that linear approximation is appropriate while also wide enough to produce precise estimates of the causal effect.

Increasing the bandwidth tends to increase bias from misspecification error. In other words, if the relationship between the outcome variable and the running variable is not linear, then as the bandwidth gets larger, the linear model might not fit the data well. On the other hand, reducing the bandwidth will tend to increase the variance of estimated RDD coefficients. This is because, as the bandwidth gets smaller, there are fewer observations that can be used to estimate the model (a smaller effective sample size). This trade-off is often called the bias-variance trade-off.

[12] This includes using the Akaike Information Criterion or the Bayesian Information Criterion for model selection (Calonico et al. 2014; Lee and Lemieux 2010).

Several procedures exist for optimizing the bias-variance trade-off to identify the optimal bandwidth for a given regression function. Two of the popular approaches include (1) cross-validation (Calonico et al. 2014; Imbens and Kalyanaraman 2012), and (2) the plug-in method (Imbens and Kalyanaraman 2012).[13]

Within the chosen bandwidth, it is common to adopt a weighting scheme that assigns weights based on where observations are located relative to the threshold. One common weighting scheme is triangular kernel weighting, which gives relatively more weight to the observations that are closer to the threshold. Other weighting schemes are often used for sensitivity analyses (see Section IV for details about sensitivity testing) such as the uniform kernel function, which gives equal weight to all observations within the bandwidth. Weighting functions can also be easily implemented using the `rdrobust` package in Stata and the `crs` packages in R.[14]

D. Choosing the functional form and bandwidth simultaneously

Evaluators often choose the best bandwidth for a given regression function. However, a different function (and associated bandwidth) might better fit the data. For example, if there are few observations around the threshold and the relationship between the outcome and the running variable is well described using a quadratic or cubic regression function, a local linear regression and associated bandwidth might perform poorly. This is because the bandwidth selection method will choose a bandwidth in which the data can be approximated by a linear regression. For outcome variables that have a complex relationship with the running variable, this bandwidth will tend to be small. A small bandwidth will have fewer data points and might therefore lead to noisy estimates.

As an alternative, evaluators can choose regression functions and bandwidths simultaneously (Hall and Racine 2015; Pei et al. 2021). The idea is to find the best bandwidth for local regression functions with different complexity and choose the combination of regression function and bandwidth that best fits the data. The `crs` packages in R provides an automated way of selecting the best functional form and associated bandwidth.

E. Confidence intervals for valid inference

Evaluators should accurately characterize the confidence intervals around RDD estimates obtained using the local nonparametric approach. The recommended bandwidth selection procedures described in the previous section are designed to choose a bandwidth that optimizes the bias-variance trade-off; they allow for some bias if the reduction in variance is large enough. Therefore, using conventional confidence intervals—which assume no bias or misspecification error—is inappropriate and could result in confidence intervals that do not capture the true impacts of the demonstration 95 percent of the time. In most cases, a conventional confidence interval that is not adjusted to account for bias will almost certainly be too small (Cattaneo et al. 2020b). Robust bias-corrected confidence intervals should be used and are available as the default option in the `rdrobust` software package mentioned earlier.

[13] Both methods are available in the user-written `rdrobust` program for Stata, R, and Python (https://rdpackages.github.io/) and are designed to choose a bandwidth that minimizes mean-square error. Both approaches are iterative and computationally intense. Jacob et al. (2012) provide a detailed description of each method should evaluators wish to write programs to choose the optimal bandwidth.

[14] The crs packages in R are available at: https://cran.r-project.org/package=crs.

IV. Sensitivity Analyses

Sensitivity analyses examine how changes in the assumptions of a model affect the results. While implementing an RDD, evaluators must make several informed design choices, such as which regression function to use and how to weight observations near the cutoff (see Section III). Even when following best practices, evaluators must understand how reasonable design choices affect estimates of demonstration impacts obtained from an RDD. If results do not vary much or are not sensitive to changes in assumptions researchers can be more confident that estimated impacts are valid. When using local nonparametric estimation, evaluators should consider assessing the sensitivity of estimates to changes in the following parameters:

- **Polynomial order of the regression function.** If a local linear regression function is used, evaluators should consider assessing the sensitivity of estimates to higher polynomial orders such as local quadratic regressions. A new bandwidth should be chosen for the alternative regression function using the methods described in Section III.
- **Weighting scheme.** Different kernel functions specify how observations near the cutoff should be weighted relative to those further away. Evaluators should consider alternative kernel weights, in sensitivity analyses.

In addition, evaluators can consider estimating global parametric models in sensitivity analyses. Although local nonparametric regressions are generally preferred over global parametric models (see Section III.B), the latter can provide a useful sensitivity check on estimates of demonstration impacts. When estimating a global parametric model, evaluators can consider linear or a higher-order polynomial specification (such as quadratic or cubic) if this fits the data better.

Finally, evaluators can assess whether results are sensitive to the inclusion of controls, such as beneficiary demographics or health care needs. If estimates are sensitive to removing or adding covariates to the model, then those who meet the threshold might have different characteristics than those who do not, suggesting that RDD assumptions are less plausible (Calonico et al. 2019).[15]

[15] If being just above or just below the threshold is as good as random, then including covariates can increase the precision of RDD estimates.

Sensitivity analyses: Empirical example

Wallace et al. (2021) investigated the association of Medicare with racial and ethnic disparities in health outcomes and access to care. The authors used an RDD to compare outcomes by race before and after 65 years, the age at which Medicare eligibility begins. They used data from the Behavioral Risk Factor Surveillance System and data from the Centers for Disease Control and Prevention's Wide-Ranging Online Data for Epidemiologic Research to study the impact of Medicare eligibility on insurance coverage, self-reported access to a usual source of care, cost-related barriers to care, flu vaccination rates, and other outcomes for each racial and ethnic group. They found that eligibility for Medicare at age 65 was associated with reductions in racial and ethnic disparities in access to care, insurance coverage, and self-reported health.

The main model is a local linear regression with a uniform kernel. The authors used a data-driven method to select the bandwidth that optimizes the bias-variance trade-off.

In sensitivity analyses, the authors re-estimate the model using a triangular kernel, which places more weight on observations closer to the threshold (unlike the uniform kernel which gives equal weight to all observations). They also estimate parametric regression discontinuity models with linear or quadratic age trends, with and without adjusting for covariates. The results from the sensitivity tests were qualitatively the same as in the main analysis, providing increased confidence in the results.

V. Robustness Checks

If the RDD assumptions described in Section II do not hold, estimates from an RDD will not represent the causal impact of a section 1115 demonstration on the outcome of interest. This section describes formal robustness checks to assess possible violations of RDD assumptions. These robustness checks include (1) checking for balance on covariates and placebo outcomes near the threshold, (2) investigating the density of the running variable, and (3) using artificial thresholds (or placebo thresholds). Evaluators using an RDD should conduct all three robustness checks to increase confidence in evaluation findings.

A. Checking for balance on covariates and placebo outcomes near the threshold

This analysis checks for differences or discontinuities in observed beneficiary characteristics just below and just above the threshold. In other words, it checks for evidence against balance on beneficiary characteristics near the threshold. Evaluators should consider checking for discontinuities in two types of variables: predetermined covariates and placebo outcomes.

Predetermined covariates are variables determined before beneficiaries enroll in the demonstration (Cattaneo et al. 2020b). For example, demographic characteristics and health care use in the pre-demonstration (baseline) period are predetermined covariates. Income also falls into this category if it is not used to define the eligibility threshold. Dissimilar beneficiary characteristics above and below the threshold suggest that the following two assumptions (assumptions 1 and 2 from Section II.A) do not hold: (1) people have no control of whether they are above or below the threshold, and (2) characteristics affecting the outcome (other than eligibility) are continuous across the threshold.

Placebo outcomes are post-treatment variables that evaluators do not expect the demonstration to impact (Cattaneo et al. 2020b). The choice of a placebo outcome depends on the demonstration type, policies and programs within the demonstration, and outcome of interest. For example, screenings for diabetes would be a reasonable placebo outcome for an evaluation studying the impact of an adult dental pilot program for dually-eligible adults, but would not be a good placebo outcome for evaluating a diabetes prevention program for high-risk adults. Additionally, in a demonstration program that provides primary care services to the newly eligible beneficiaries, a good placebo outcome could be hospitalizations for accidents or injuries- an outcome that is less likely to be affected by primary care use. Discontinuities in placebo outcomes might indicate changes at the threshold unrelated to the demonstration that might impact the outcome of interest. This violates the assumption that characteristics affecting the outcome (other than eligibility) are continuous across the threshold.

To formally test for discontinuities in covariates and placebo outcomes, evaluators should implement a standard RDD following the best practices described in Section III, using the covariates or placebo outcomes in place of the outcome of interest. The optimal functional form, weights, and the bandwidths might differ for covariates and placebo outcomes (Cattaneo et al. 2020b).[16]

B. Investigating the density of the running variable

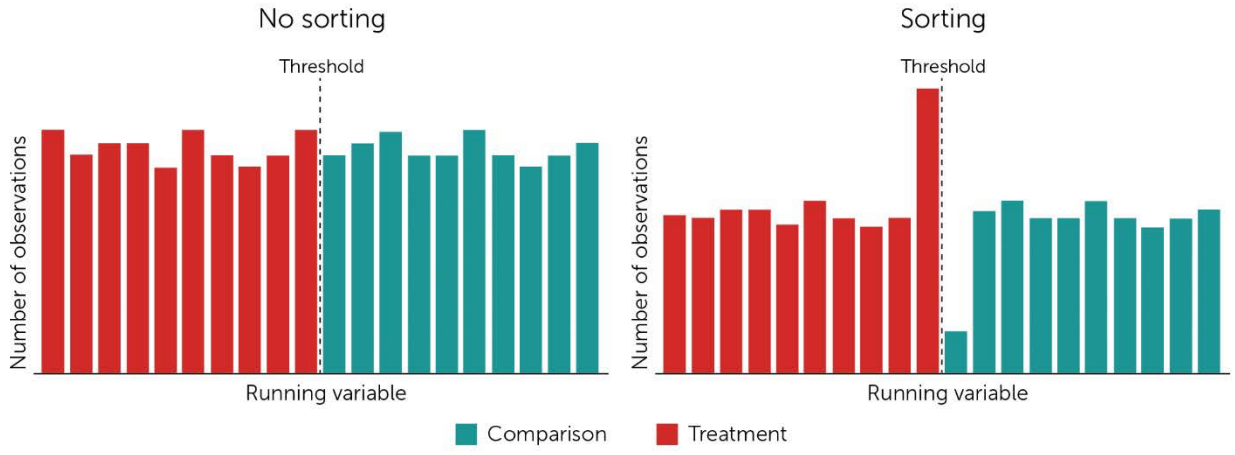
This analysis checks whether there is a roughly similar number of observations on either side of the threshold. If beneficiaries are able to sort across the threshold—for example, in an income-based

[16] See also "Matching Methods for the Evaluation of Section 1115 Demonstrations" (Pohl et al. 2023) available at: https://www.medicaid.gov/medicaid/section-1115-demo/downloads/evaluation-reports/matching-methods.pdf for how to check for covariate balance.

demonstration, by reducing their income to qualify for premium assistance—evaluators would observe that there are more observations that meet the threshold than those that do not. As Section II describes, the ability to lower income might be correlated with certain demographic characteristics, observed or unobserved, thereby making the group just below the threshold not comparable to the group just above. Consequently, RDD results might not be a valid estimate of the causal effect. Plotting the density of the running variable provides information about whether beneficiaries have control over whether they are just above or just below the threshold.

Evaluators can assess the number of observations by plotting the density of the running variable around the eligibility threshold and checking for a visible change at the threshold. Figure V.1 shows two examples of density plots of the number of observations against the running variable. In the left panel, the density is smooth across the threshold, indicating no evidence of sorting. In the right panel, there is an abrupt change in density across the threshold, with fewer observations above the threshold and more below. This suggests that beneficiaries who were just above the threshold were able to place themselves right below by adjusting the running variable (for example, income). Evaluators can also conduct a formal density test to check if density of the running variable is continuous across the threshold. Local polynomial density tests follow the RDD framework and are considered a best practice because of their easy implementation (Cattaneo et al. 2020a).[17]

Figure V.1. Density of the running variable and implication on sorting



Source: Example based on Cattaneo et al. (2020b); data points are illustrative only and not based on actual data.

C. Using artificial thresholds (or placebo thresholds)

This analysis examines whether trends in the outcome variable change abruptly at points other than the threshold, as a check of the third assumption Section II describes. This assumption requires that the outcome variable is not discontinuous at other points close to the threshold. For example, evaluators might want to estimate the impact of a section 1115 demonstration in which beneficiaries with incomes below 133 percent of the FPL qualify for premium assistance. In addition to estimating the demonstration impact at 133 percent of the FPL, evaluators should check whether the outcome is discontinuous at

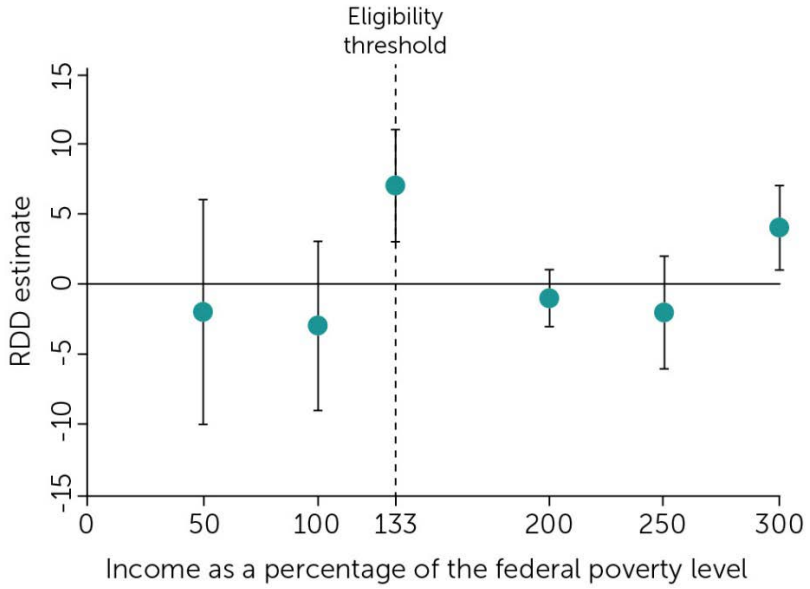
[17] The required input is the running variable, and the output shows whether the null hypothesis of no significant change in density can be rejected. Local polynomial density tests are included in the `rddensity` package for Stata and R (https://rdpackages.github.io/rddensity/).

incomes other than 133 percent of the FPL. If discontinuities are present close to the artificial or placebo thresholds, it is more difficult to attribute the discontinuity at the eligibility threshold to demonstration participation.

When conducting this test, evaluators should consider a range of artificial thresholds above and below the real threshold. Visually inspecting the outcome against the running variable can inform the set of artificial thresholds; values that show jumps in the outcome are potential artificial thresholds. Once they select artificial thresholds, evaluators should implement a standard RDD using the artificial threshold. If the estimated impact at an artificial or placebo threshold is statistically significant, this would indicate a violation of the third assumption in Section II.

In the earlier example of an income-based section 1115 demonstration, evaluators could estimate RDD effects at a range of thresholds below and above 133 percent of the FPL and compare them to the estimated effect at 133 percent of the FPL. Figure V.2 illustrates a hypothetical test result in which the RDD estimate is statistically significant at the real threshold (x = 133 percent of the FPL) and the estimate is not statistically different than zero at the artificial thresholds, except at x = 300 percent of the FPL. At incomes close to the actual threshold (133 percent of the FPL), no discontinuities occur at incomes other than the threshold, which provides support for implementing RDD at the threshold. Although a discontinuity is present at the 300 percent of the FPL, it is far from the actual threshold and thus implementing RDD at the threshold might still be appropriate. Evaluators should combine evidence of discontinuities with additional available evidence to inform their assessment of the appropriateness of an RDD.

Figure V.2. RDD estimation for real and artificial thresholds



Source: Example based on Cattaneo et al. (2020b); data points are illustrative only and not based on actual data.
RDD = regression discontinuity design.

Robustness checks: Empirical example

Wherry et al. (2018) investigated the long-term impact of childhood Medicaid coverage on health care utilization. To study this question, the authors took advantage of discrete changes in years of childhood Medicaid eligibility on September 30, 1983, the cutoff date specified in many expansions of Medicaid for children of low-income families in the late 1980s and early 1990s. In particular, they used RDD to compare hospitalizations during the adulthood of beneficiaries just below and above the birthdate threshold; these beneficiaries had similar age but different years of Medicaid eligibility. They found that for Blacks—who were disproportionally affected by Medicaid expansions—approximately five additional years of Medicaid eligibility in childhood reduced hospitalizations in adulthood by 7 to 15 percent.

The two main models were: (1) a global polynomial regression and (2) a local linear regression with a triangular kernel and data-driven optimal bandwidths. In robustness checks, the authors tested for discontinuities in beneficiary characteristics across the threshold. One concern was that beneficiaries who would gain Medicaid eligibility move out of state in order to receive additional coverage, making beneficiaries above the threshold not comparable to those below. To investigate this, Wherry et al. (2018) tested for discontinuities in the fraction of beneficiaries that resided in a state that was different from the state of birth at age 25. They found no evidence of discontinuities near the threshold, indicating that beneficiaries who gained childhood Medicaid eligibility were no more likely to migrate to a different state compared to those who did not gain eligibility. They also checked for discontinuities in a placebo outcome—hospitalizations for appendicitis and injuries, acute conditions that are less likely to be influenced by coverage in childhood and found no evidence of an effect near the threshold. Finally, they tested for discontinuities at placebo birthdate thresholds in the sample of Black beneficiaries and found little to no statistically significant discontinuities in hospitalizations at the artificial thresholds.

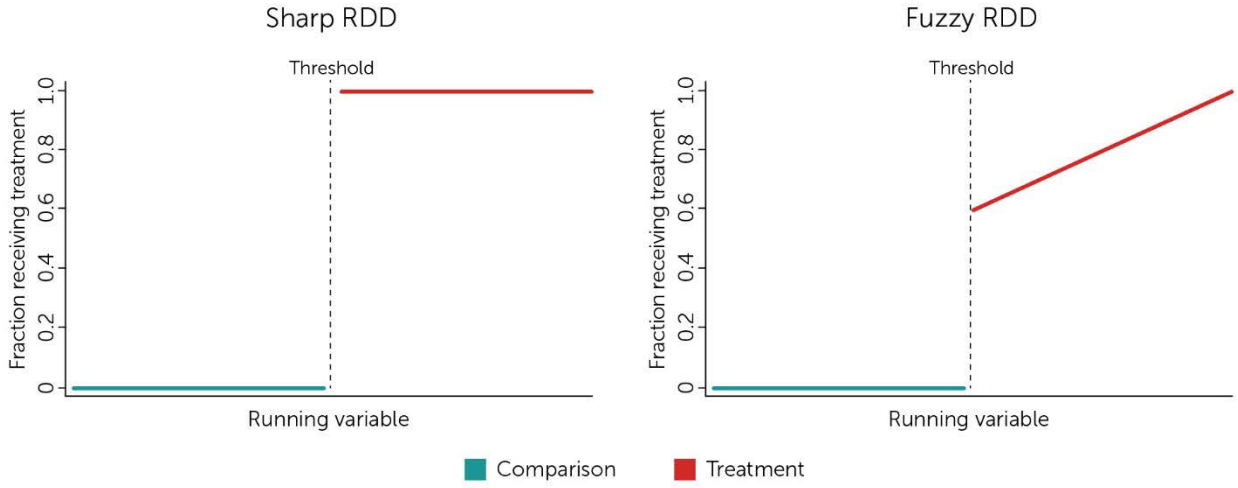
VI. Extensions

A. Fuzzy RDD

Sharp RDD assumes that all eligible beneficiaries participate in the demonstration and those not eligible do not participate in the demonstration. However, there might be settings where this is not the case—that is, some eligible beneficiaries do not participate in the demonstration or some beneficiaries who do not meet the eligibility threshold do participate. For example, it may be difficult to reach eligible beneficiary populations or increase beneficiary awareness of services. This may have been especially true during the COVID-19 PHE when beneficiary populations may have been hesitant to engage with the health care system. This is generally known as imperfect compliance to treatment. For example, a section 1115 demonstration may provide intensive care coordination for individuals with health risk scores above a certain threshold. Though the state and health plans may try to engage all beneficiaries whose health risk scores are above the threshold, some beneficiaries may not participate in the intensive care coordination program. They may not be aware they are eligible or may not want to be part of the demonstration. An evaluator interested in the impact of the program on health outcomes can implement a fuzzy RDD to investigate the impact of the intensive care coordination program.

In both sharp and fuzzy RDDs, all beneficiaries are eligible to participate in the demonstration (or a particular demonstration component) above the eligibility threshold. However, unlike sharp RDD, because compliance is not perfect, not all beneficiaries actually participate in the demonstration (for example, receive intensive care coordination) above the threshold in fuzzy RDD. Figure VI.1 illustrates the difference between sharp and fuzzy RDDs.[18]

Figure VI.1. Sharp and fuzzy RDD



Source: Data points are illustrative only and not based on actual data.

RDD = regression discontinuity design.

[18] The right panel of Figure VI.1 illustrates one case of a fuzzy RDD in which compliance to eligibility status is perfect below the threshold but is imperfect above the threshold. That is, all beneficiaries below the threshold do not participate in the demonstration, however, some beneficiaries above the threshold do not participate in the demonstration. This case is referred to as one-sided compliance. Fuzzy RDD can also exhibit two-sided noncompliance, where some beneficiaries below the threshold also participate in the demonstration policy of interest.

Eligibility for a demonstration component, therefore, differs from take-up of the demonstration component in fuzzy RDD. Evaluators are often interested in both (1) the effect of being eligible for the demonstration component and (2) the effect of participating in the demonstration component. Using the example above, evaluators are interested in (1) the effect of being eligible for the intensive care coordination program and (2) the effect of the intensive care coordination program itself. To determine the effect of being eligible for the program (often called the intention-to-treat effect), evaluators can follow the standard analyses for sharp RDD.

Estimating the effect of participating in the program requires that standard RDD assumptions hold and that there are no defiers in the study sample. In this section 1115 example, defiers are beneficiaries who participate in the program if they miss the threshold and who do not participate in the program if they meet the threshold. This assumption is satisfied when the demonstration sample contains only three groups of beneficiaries: (1) beneficiaries who participate in the program if they have a health risk score above the eligibility threshold and do not participate if they are below, (2) beneficiaries who never participate in the program regardless of whether they are above or below the threshold, and (3) beneficiaries who always participate in the program regardless of whether they are above or below the threshold. This is often referred to as the monotonicity assumption.

If these assumptions hold, evaluators can estimate the causal effect of the intensive care coordination program on the health outcomes of interest. In general, the causal impact is equal to the effect of meeting the threshold on the outcome (the intention-to-treat effect), divided by the effect of meeting the threshold on receiving intensive care coordination services (often called the first stage).[19] Like the sharp RDD estimate, the fuzzy RDD estimate is specific to the group of observations close to the threshold.

Implementing the fuzzy RDD is analogous to implementing another type of causal analytic approach, called the instrumental variables (IV) analysis—for example, in a randomized controlled trial with noncompliance (Angrist and Imbens 1994; Imbens and Wooldridge 2009). As in the case for IV analysis, fuzzy RDD treatment effects are estimated for compliers, that is, beneficiaries who participate in the demonstration if they are eligible and who do not participate if they are not eligible. Because the decision to comply is often correlated with the impact estimate, fuzzy RDD estimates might not apply to all beneficiaries close to the threshold. For example, beneficiaries just above the threshold who participate in the intensive care coordination program may take a more active role in their health care in general. Impacts estimated among these beneficiaries might not generalize to all beneficiaries close to the threshold, much less the demonstration population as a whole. IV and fuzzy RDD estimates are sometimes called the local average treatment effect, reflecting the local nature of the estimates.

B. Local randomization approach to RDD

RDD takes advantage of the idea that in settings where participation in a section 1115 demonstration is based on a continuous variable, beneficiaries close to but on different sides of the threshold are likely similar. RDD does not require the absence of a correlation between the outcome measure and the running variable, such as age or income (De la Cuesta and Imai, 2016). Instead, evaluators estimate the relationship between the running variable and the outcome measure on either side of the threshold. If all

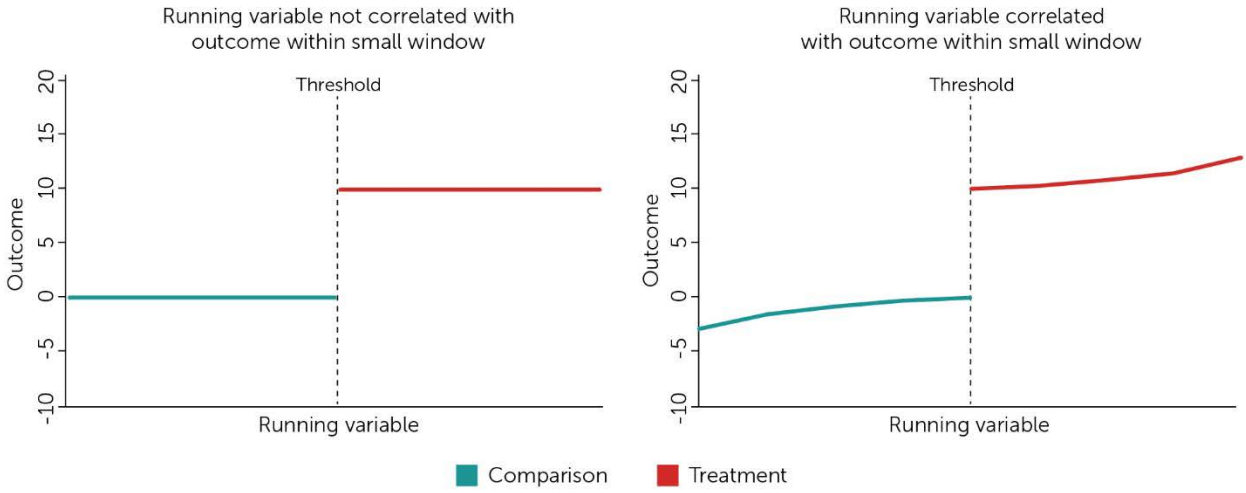
[19] Evaluators should use RDD or instrumental variable packages in R and Stata to estimate the fuzzy RDD (rather than manually dividing the two estimates) to ensure that the standard errors are correct.

the RDD assumptions hold, a discontinuity (or jump) in the outcome trend at the threshold is attributed to participating in the demonstration.

There might be settings, however, where the running variable and the outcome measure are not correlated within a small window near the threshold. In settings like this, evaluators can use an alternative approach called the local randomization approach. This approach is particularly helpful when the running variable is discrete and takes on a limited set of values—for example, in a section 1115 demonstration that applies to individuals above an age threshold, but where evaluators can only observe year of birth rather than exact date of birth. In this scenario, because the running variable does not take on values close enough to the threshold, the standard continuity-based approach to RDD might not perform well (Lee and Card 2008). However, if there is a small window around the age threshold where treatment is *as good as randomly assigned*, then the local randomization approach provides valid impact estimates.[20] [21]

Figure VI.2 illustrates a small window near the threshold in which the running variable is not related to the outcome and treatment is as good as randomly assigned (left panel), and a scenario in which the running variable is correlated with the outcome; that is, the outcome is increasing in the running variable (right panel).

Figure VI.2. Correlation between running variable and outcome measure within a small window near the threshold



Source: Data points are illustrative only and not based on actual data.

Under a local randomization RDD, evaluators analyze the data as though it were a randomized controlled experiment by comparing mean outcomes for individuals just below and just above the threshold. If the

[20] Note that the continuity-based approach requires that treatment be as good as randomly assigned at the cut off, while the local randomization approach requires that treatment be as good as randomly assigned within a small window around the cut off.

[21] An alternative approach to implementing RDD when the running variable is discrete or takes on a moderate number of values is to adjust the confidence intervals to account for the additional risk of model misspecification in these settings. This risk arises when there are few observations near the threshold and evaluators are forced to choose a bandwidth that is too wide (Kolesár and Rothe 2018; Armstrong and Kolesár 2020). Evaluators can use the R package `rdhonest` to appropriately adjust confidence intervals when the running variable is discrete (https://github.com/kolesarm/RDHonest).

number of observations inside the bandwidth is large, evaluators can use the standard set of statistical tests based on large sample limiting distributions (for example, t-tests for differences in means).[22] Evaluators should conduct robustness checks and sensitivity analyses to shed light on whether the demonstration context satisfies assumptions for local randomization.[23]

In general, when the running variable is continuous, the local randomization approach requires stronger assumptions than the continuity-based approach. In these cases, we recommend evaluators use the standard continuity-based approach for the main RDD analysis and use the local randomization approach as a sensitivity check. However, in settings where the running variable is discrete (only a few unique values), the local randomization method might be the only valid method of estimating impacts (Cattaneo and Titiunik 2022).[24]

[22] If the number of observations within the bandwidth is small, evaluators should use finite sample methods such as the Fisherian inference approach. For further reading, see Cattaneo and Titiunik (2022).

[23] An alternative approach to implementing RDD when the running variable is discrete or takes on a moderate number of values is to adjust the confidence intervals to account for the additional risk of model misspecification in these settings. This risk arises when there are few observations near the threshold and evaluators are forced to choose a bandwidth that is too wide (Kolesár and Rothe 2018; Armstrong and Kolesár 2020). Evaluators can use the R package `rdhonest` to appropriately adjust confidence intervals when the running variable is discrete (https://github.com/kolesarm/RDHonest).

[24] For further reading on the local randomization approach to RDD, see Sekhon and Titiunik (2017), Branson and Mealli (2019) and Cattaneo and Titiunik (2022). For further reading on implementing RDD with discrete running variables, see Lee and Card (2008), Kolesár and Rothe (2018), and Cattaneo and Titiunik (2022).

VII. Conclusion

RDD can be a powerful empirical approach for evaluating section 1115 demonstrations because demonstrations often involve eligibility thresholds and because RDD does not require pre-intervention time-series data or data from a comparison group not affected by the demonstration. Evaluators should include adequate level of details on the applicability, feasibility and implementation of the approach in the demonstration evaluation design and evaluation reports, as appropriate.

Key points to specify in evaluation designs. When using an RDD approach in the evaluation design for a section 1115 demonstration, evaluators should specify the running variable (including the eligibility threshold and the range of the running variable in the available data), outcome variables, and sensitivity and robustness checks to be conducted. Evaluation designs should also include a discussion of the RDD assumptions and whether they are likely to be met in the section 1115 demonstration context.

Key points to document in evaluation reports. In section 1115 demonstration evaluation reports, evaluators should provide sufficient detail such that readers can understand the design choices made and whether any RDD assumptions required for valid estimates are violated. The report should specify the functional form of the model, how the bandwidth was chosen and whether a weighting scheme was used. In addition, the evaluation report should describe sensitivity tests and robustness checks conducted. One of the main challenges to the appropriateness to RDD relates to the possibility of manipulation of the running variable. Evaluators should discuss, based on their robustness analysis, whether it is reasonable to assume that beneficiaries right above and right below the eligibility threshold are likely to be similar in all characteristics other than participation in the demonstration and any demonstration outcomes. Finally, evaluation reports should include a discussion of the external validity of RDD estimates.

Overall, the RDD approach is particularly helpful in settings where these comparator groups are not available, or where other policy changes that affect the treatment group during the pre-intervention period make time-series approaches such as interrupted time series less appropriate. When using RDD, evaluators should be aware of the local nature of RDD estimates; causal effects estimated among beneficiaries close to the threshold might not generalize to the demonstration population as a whole (lack of external validity). Notwithstanding, RDD estimates can be helpful and informative for policymakers seeking to understand the impact of expanding eligibility near the threshold.

References

Armstrong, Timothy B., and Michal Kolesár. "Simple and Honest Confidence Intervals in Nonparametric Regression." *Quantitative Economics,* vol. 11, no. 1, 2020, pp. 1–39.

Angrist, Joshua, and Guido Imbens. "Identification and Estimation of Local Average Treatment Effects." *Econometrica,* vol. 62., no. 2, 1994, pp. 467–475.

Bradley, Katharine, Jessica Heeringa, R. Vincent Pohl, James D. Reschovsky, and Maggie Samra. "Selecting the Best Comparison Group and Evaluation Design: A Guidance Document for State Section 1115 Demonstration Evaluations." Washington, DC: Mathematica, October 2020.

Branson, Zach, and Fabrizia Mealli. "The Local Randomization Framework for Regression Discontinuity Designs: A Review and Some Extensions." Preprint. November 2019.

Calonico, Sebastian, Matias D. Cattaneo, Max H. Farrell, and Rocío Titiunik. "Regression Discontinuity Designs Using Covariates." *Review of Economics and Statistics,* vol. 101, no. 3, 2019, pp. 442–451.

Calonico, Sebastian, Matias D. Cattaneo, and Rocio Titiunik. "Robust Nonparametric Confidence Intervals for Regression-Discontinuity Designs." *Econometrica,* vol. 82, no. 6, 2014, pp. 2295–2326.

Card, David, and Lara D. Shore-Sheppard. "Using Discontinuous Eligibility Rules to Identify the Effects of the Federal Medicaid Expansions on Low-Income Children." *Review of Economics and Statistics,* vol. 86, no. 3, 2004, pp. 752–766.

Cattaneo, Matias D., Michael Jansson, and Xinwei Ma. "Simple Local Polynomial Density Estimators." *Journal of the American Statistical Association,* vol. 115, no. 531, 2020a, pp. 1449–1455.

Cattaneo, Matias D., Nicolás Idrobo, and Rocío Titiunik. "A Practical Introduction to Regression Discontinuity Designs: Foundations." In *Elements: Quantitative and Computational Methods for Social Science,* edited by R. Michael Alvarez and Nathaniel Beck. Cambridge, UK: Cambridge University Press, 2020b.

Cattaneo, Matias D., and Rocío Titiunik. "Regression Discontinuity Designs." *Annual Review of Economics,* vol. 14, 2022, pp. 821–851.

De la Cuesta, Brandon, and Kosuke Imai. "Misunderstandings About the Regression Discontinuity Design in the Study of Close Elections." *Annual Review of Political Science,* vol. 19, 2016, pp. 375–396.

Finkelstein, Amy, Nathaniel Hendren, and Mark Shepard. "Subsidizing Health Insurance for Low-Income Adults: Evidence from Massachusetts." *American Economic Review,* vol. 109, no. 4, 2019, pp. 1530–1567.

Gelman, Andrew, and Guido Imbens. "Why High-Order Polynomials Should Not Be Used in Regression Discontinuity Designs." *Journal of Business & Economic Statistics,* vol. 37, no. 3, 2019, pp. 447–456.

Hall, Peter G., and Jeffrey S. Racine. "Infinite Order Cross-Validated Local Polynomial Regression." *Journal of Econometrics,* vol. 185, no. 2, 2015, pp. 510–525.

Imbens, Guido W., and Jeffrey M. Wooldridge. "Recent Developments in the Econometrics of Program Evaluation." *Journal of Economic Literature,* vol. 47, no. 1, 2009, pp. 5–86.

Imbens, Guido, and Karthik Kalyanaraman. "Optimal Bandwidth Choice for the Regression Discontinuity Estimator." *The Review of Economic Studies,* vol. 79, no. 3, 2012, pp. 933–959.

Jacob, Robin, Pei Zhu, Marie-Andree Somers, and Harold Bloom. "A Practical Guide to Regression Discontinuity." New York: MDRC, 2012. Available at http://www.mdrc.org/sites/default/files/regression_discontinuity_full.pdf.

Kolesár, Michal, and Christoph Rothe. "Inference in Regression Discontinuity Designs with a Discrete Running Variable." *American Economic Review,* vol 108, no. 8, 2018, pp. 2277–2304.

Lee, David S., and David Card. "Regression Discontinuity Inference with Specification Error." *Journal of Econometrics,* vol. 142, no. 2, 2008, pp. 655–674.

Lee, David S., and Thomas Lemieux. "Regression Discontinuity Designs in Economics." *Journal of Economic Literature,* vol. 48, no. 2, 2010, pp. 281–355.

Pei, Zhuan, David S. Lee, David Card, and Andrea Weber. "Local Polynomial Order in Regression Discontinuity Designs." *Journal of Business & Economic Statistics,* 2021, pp. 1–9.

Pohl, R. Vincent, and Katharine Bradley. "Selection of Out-of-State Comparison Groups and the Synthetic Control Method." Washington, DC: Mathematica, October 2020.

Pohl, R. Vincent, Lianlian Lei, and Matthew Niedzwiecki. "Matching Methods for the Evaluation of Section 1115 Demonstrations." Washington, DC: Mathematica, November 2022. Available at: https://www.medicaid.gov/medicaid/section-1115-demo/downloads/evaluation-reports/matching-methods.pdf.

Sekhon, Jasjeet S., and Rocío Titiunik. "On Interpreting the Regression Discontinuity Design as a Local Experiment." *Regression Discontinuity Designs,* vol. 38, 2017, pp. 1–28.

Wallace, Jacob, Karen Jiang, Paul Goldsmith-Pinkham, and Zirui Song. "Changes in Racial and Ethnic Disparities in Access to Care and Health Among US Adults at Age 65 Years." *JAMA Internal Medicine,* 2021.

Wherry, Laura R., Sarah Miller, Robert Kaestner, and Bruce D. Meyer. "Childhood Medicaid Coverage and Later-Life Health Care Utilization." *The Review of Economics and Statistics*, vol. 100, no. 2, 2018, pp. 287–302.

www.mathematica.org

Improving public well-being by conducting high quality, objective research and data collection

PRINCETON, NJ ■ ANN ARBOR, MI ■ CAMBRIDGE, MA ■ CHICAGO, IL ■ OAKLAND, CA ■ SEATTLE, WA ■ TUCSON, AZ ■ WASHINGTON, DC ■ WOODLAWN, MD

Centers for Medicare & Medicaid Services

7500 Security Boulevard Baltimore, MD 21244