

The Regression Discontinuity Design in Medical Science

Matias D. Cattaneo¹ and Rocío Titiunik²

¹Department of Operations Research and Financial Engineering, Princeton University.

²Department of Politics, Princeton University.

August 4, 2025

The Regression Discontinuity (RD) design is a research strategy to estimate the causal effect of a treatment on an outcome. Its defining feature is that the treatment is assigned according to a threshold rule: all units receive a score (also known as a running variable), and the treatment is assigned to those units whose score is equal to or above a known cutoff, but not assigned to those units whose score is below the cutoff. In the sharp RD design, compliance with treatment assignment is perfect. In the fuzzy RD design, compliance with the assignment is imperfect, so that some units with score below the cutoff may nonetheless receive the treatment, while some units with score at least equal to the cutoff may not receive it. Generalizations include multi-cutoff and multi-score (or geographic) RD designs. See [Cattaneo and Titiunik \(2022\)](#) for a literature review.

[Goulden et al. \(2021\)](#) use the fuzzy RD design to study the effect of intravenous radiocontrast exposure on kidney function for patients who underwent D-dimer testing because of suspected pulmonary embolism. Medical guidelines recommend performing a computed tomographic pulmonary angiogram (CTPA), which uses contrast, whenever the D-dimer score is above 500 ng/mL. The authors compare patients with D-dimer score above and below that cutoff to estimate the effect of CTPA on the glomerular filtration rate (eGFR), a measure of long-term kidney function. While compliance with the medical recommendation was not perfect, the authors show that patients above the cutoff were much more likely to receive CTPA than patients below it. Medical guidelines are a common source of RD designs in the medical sciences ([Cattaneo et al., 2023](#)).

1 Why is the Identification Strategy Used?

When random assignment of interventions is not possible, the study of causal effects must rely on observational studies where the comparability of units assigned to control and units assigned to treatment is not guaranteed ([Rosenbaum, 2010](#)). If unobserved confounders systematically differ between these two groups, it is difficult to interpret their outcome differences as a causal effect of the treatment. In the CTPA study, patients assigned to CTPA were on average older and have

higher general morbidity than patients not assigned to CTPA, and most likely they also differed in unobserved characteristics. Therefore, a simple comparison of kidney function between these two groups would likely be invalid. The RD design offers a strategy to overcome this challenge.

2 How is the identification Strategy Used?

The RD design removes the effect of unobserved confounders by focusing only on units whose scores are close to the cutoff. Intuitively, this “localization” is based on the idea that units with very similar scores will be ex-ante comparable regardless of their treatment assignment. RD designs are often viewed as “a much closer cousin of randomized experiments than other competing methods” ([Lee and Lemieux, 2010](#), p. 289).

What causal effects (estimands) is the strategy used for?

In the sharp RD design, the parameter of interest is the average causal effect of the treatment on the outcome for units whose scores are close to the cutoff ([Cattaneo et al., 2020](#)). In the fuzzy RD design, there are three parameters of interest, all defined for units whose scores are near the cutoff. The first is the effect of the treatment assignment on the treatment received, known as the first-stage (or take-up) average effect. The second is the effect of the treatment assignment on the outcome, known as the intention-to-treat average effect. The third is the effect of the treatment received on the outcome for those units who were induced to receive the treatment because their score exceeded the cutoff, typically known as the complier average treatment effect. The latter effect requires stronger assumptions for causal interpretation ([Cattaneo et al., 2024](#), Section 3).

In the CTPA study, the first stage effect captures how much more likely patients with D-dimer score just above the cutoff are to receive CTPA than patients just below it. The intention-to-treat effect captures how the average eGFR of patients with D-dimer score just above the cutoff differs from the average eGFR of patients just below it, regardless of how many of those patients actually received CTPA. Finally, the complier treatment effect captures the effect of actually receiving CTPA on eGFR, for those patients with D-dimer score above 500 who would not have received CTPA otherwise.

What causal assumptions does the strategy depend on?

The RD design is based on a fundamental assumption: units with score near the cutoff have, on average, the same observable and unobservable baseline characteristics regardless of whether they are assigned to control or treatment. In the CTPA study, this assumption requires that patients whose D-dimer score is barely below 500 are ex-ante comparable to patients whose D-dimer score is barely above 500.

This identifying assumption can be formally justified in two ways. The continuity-based framework postulates that the units’ average potential outcomes given the score are continuous near the cutoff. This framework relies on large sample approximations and extrapolation, employing

information of units with score near the cutoff to infer causal effects at the cutoff (Cattaneo et al., 2020). The local-randomization framework postulates that units with score within a small region near the cutoff behave as-if randomly assigned to control and treatment. This framework relies on the stronger assumption of having an experimental design among units with score near the cutoff, and employs tools from the analysis of experiments (Cattaneo et al., 2024, Section 2).

The fundamental RD identifying assumption may fail if units can manipulate their score. In the CTPA study, if patients with pre-existing conditions are more worried about developing pulmonary embolism, and know the treatment assignment rule, they may try to change their D-dimer score to be above 500 and thus be eligible for treatment. While in this example altering the score is not feasible because patients cannot manipulate laboratory measurements, in other settings patients may behave strategically (e.g., when self-reporting medical information). This so-called “sorting around the cutoff” may destroy the comparability of patients with scores near the cutoff.

How are causal effects estimated using the data?

Researchers will need data on the outcome and the score, and knowledge of the cutoff value and whether the treatment assignment had perfect (sharp) or imperfect (fuzzy) compliance. The average treatment effects are estimated by first keeping only observations with score in a neighborhood around the cutoff, and then conducting estimation and inference as determined by the framework used. The continuity-based framework employs local linear regression methods, tuned to minimize mean square error, and robust bias-correction for uncertainty quantification (Cattaneo et al., 2020, and Cattaneo et al., 2024, Section 2). The local-randomization framework employs methods for the analysis of experiments (Cattaneo et al., 2024, Section 3).

How should the identification strategy be evaluated?

While the fundamental RD assumption of comparability near the cutoff is untestable, there are several falsification tests that researchers can employ to empirically assess its plausibility. The main strategy, akin to studying covariate balance in randomized controlled experiments, is to check for a null treatment effect on baseline covariates. In the CTPA study, the average age among all individuals assigned to receive CTPA was roughly 60 years old, while the average age among those who were not assigned to CTPA was 50 years old, a difference that can confound measures of kidney function. However, when restricting the analysis to patients with D-dimer score near the cutoff of 500, that difference in age disappeared (Goulden et al., 2021, Figure 1B). This balancing of baseline observed characteristics for units with score close to the cutoff is expected in a valid RD design.

Other falsification strategies include assessing whether the number of units with score close to the cutoff is roughly the same between groups assigned to control and treatment, varying the neighborhood around the cutoff, considering artificial cutoffs, and removing observations closest to the cutoff. See Cattaneo et al. (2020, Section 5).

3 Limitations of the Strategy

Valid RD designs estimate treatment effects that are “local” in nature, that is, only for units whose scores are close to the cutoff. These effects may not easily generalize to units whose scores are far away from the cutoff. Therefore, RD designs offer high internal validity at the expense of potentially low external validity.

4 How Should the Identification Strategy be Interpreted

Goulden et al. (2021) find that patients with D-dimer score barely below and barely above the 500 ng/mL cutoff have similar average nephrotoxicity as measured by eGFR. Under the assumption that these patients do not differ in unobserved confounders, we conclude that intravenous contrast administered for CTPA has no harmful causal effect on kidney function. This causal finding pertains to patients with D-dimer score near 500, and may be different from the causal effects for patients with lower or higher D-dimer scores.

Acknowledgments

We thank Issa Dahabreh for comments and suggestions. We are indebted to our collaborators Sebastian Calonico, Rajita Chandak, Max Farrell, Yingjie Feng, Brigham Frandsen, Nicolas Idrobo, Michael Jansson, Luke Keele, Xinwei Ma, Ricardo Masini, Kenichi Nagasawa, Filippo Palomba, Jasjeet Sekhon, Gonzalo Vazquez-Bare, and Ruiqi (Rae) Yu for their intellectual contributions to our research program on RD designs. Cattaneo and Titiunik gratefully acknowledge financial support from the National Science Foundation (SES-2019432 and SES-2241575).

References

- Cattaneo, M. D., Idrobo, N., and Titiunik, R. (2020), *A Practical Introduction to Regression Discontinuity Designs: Foundations*, New York, NY: Cambridge Elements: Quantitative and Computational Methods for Social Science, Cambridge University Press.
- (2024), *A Practical Introduction to Regression Discontinuity Designs: Extensions*, New York, NY: Cambridge Elements: Quantitative and Computational Methods for Social Science, Cambridge University Press.
- Cattaneo, M. D., Keele, L., and Titiunik, R. (2023), “A Guide to Regression Discontinuity Designs in Medical Applications,” *Statistics in Medicine*, 42, 4484–4513.
- Cattaneo, M. D., and Titiunik, R. (2022), “Regression Discontinuity Designs,” *Annual Review of Economics*, 14, 821–851.

- Goulden, R., Rowe, B. H., Abrahamowicz, M., Strumpf, E., and Tamblyn, R. (2021), “Association of intravenous radiocontrast with kidney function: a regression discontinuity analysis,” *JAMA internal medicine*, 181, 767–774.
- Lee, D. S., and Lemieux, T. (2010), “Regression Discontinuity Designs in Economics,” *Journal of Economic Literature*, 48, 281–355.
- Rosenbaum, P. R. (2010), *Design of Observational Studies*, Vol. 10, Springer.