OPEN

Regression Discontinuity Designs in Epidemiology

Causal Inference Without Randomized Trials

Jacob Bor,[a,b,c] *Ellen Moscoe,*[c] *Portia Mutevedzi,*[b] *Marie-Louise Newell,*[b,d] *and Till Bärnighausen*[b,c]

Abstract: When patients receive an intervention based on whether they score below or above some threshold value on a continuously measured random variable, the intervention will be randomly assigned for patients close to the threshold. The regression discontinuity design exploits this fact to estimate causal treatment effects. In spite of its recent proliferation in economics, the regression discontinuity design has not been widely adopted in epidemiology. We describe regression discontinuity, its implementation, and the assumptions required for causal inference. We show that regression discontinuity is generalizable to the survival and nonlinear models that are mainstays of epidemiologic analysis. We then present an application of regression discontinuity to the much-debated epidemiologic question of when to start HIV patients on antiretroviral therapy. Using data from a large South African cohort (2007–2011), we estimate the causal effect of early versus deferred treatment eligibility on mortality. Patients whose first CD4 count was just below the 200 cells/μL CD4 count threshold had a 35% lower hazard of death (hazard ratio = 0.65 [95% confidence interval = 0.45–0.94]) than patients presenting with CD4 counts just above the threshold. We close by discussing the strengths and limitations of regression discontinuity designs for epidemiology.

(*Epidemiology* 2014;25: 729–737)

Submitted 24 July 2013; accepted 07 February 2014.

From the [a]Department of Global Health, Boston University School of Public Health, Boston, MA; [b]Africa Centre for Health and Population Studies, Somkhele, South Africa; [c]Department of Global Health and Population, Harvard School of Public Health, Boston, MA; and [d]Faculty of Medicine, University of Southampton, Southampton, United Kingdom.

This research was made possible with funding from the Wellcome Trust (Africa Centre for Health and Population Studies); National Institutes of Health grants R01 HD058482-01 and 1R01MH083539-01 (T.B., M.L.N.); the Rush Foundation (J.B., T.B.); Harvard Center for Population and Development Studies (J.B.); and US Agency for International Development (USAID) Cooperative Agreement AID 674-A-12-00029 (J.B.). The contents are the responsibility of the authors and do not necessarily reflect the views of any of the funders or the US Government.

SDC Supplemental digital content is available through direct URL citations in the HTML and PDF versions of this article (www.epidem.com). This content is not peer-reviewed or copy-edited; it is the sole responsibility of the authors.

Editors' note: *A commentary on this article appears on page 738.*

Correspondence: Jacob Bor, 801 Massachusetts Avenue, Boston, MA 02118. E-mail: jbor@bu.edu. +1 617 414 1444

ISSN: 1044-3983/14/2505-0729
DOI: 10.1097/EDE.0000000000000138

Causal inference in nonexperimental studies typically requires a strong, untestable assumption: that no unobserved factors confound the relationship between the exposure and the outcome.[1] Violations of this assumption will lead to biased estimation of causal effects. The regression discontinuity design is one important quasi-experimental study design in which this assumption is not required for causal inference. Regression discontinuity designs can be implemented when the exposure of interest is assigned—at least in part—by the value of a continuously measured random variable and whether that variable lies above (or below) some threshold value. Provided that subjects cannot precisely manipulate the value of this variable, assignment of the exposure is as good as random for observations close to the threshold, and valid causal effects can be identified.[2]

The regression discontinuity design first appeared in the educational psychology literature in 1960,[3–5] was further developed in the 1970s and 1980s,[6,7] and has become well established in economics over the last 2 decades.[2,8,9] In recent years, a number of clinical and population health studies have been published in economics journals using regression discontinuity designs.[10–17] These studies have used regression discontinuity to estimate the health effects of clinical care,[10,11] health behaviors,[12,13] social determinants,[14,15] and environmental exposures[16]—questions of interest to epidemiologists. Yet regression discontinuity has not been widely adopted in epidemiology. To date, no empirical regression discontinuity studies have been published in leading epidemiology journals, and, when economics journals are excluded, just 8 such studies appear in PubMed.[17]

This paper serves as an introduction to regression discontinuity for application in epidemiology. We describe the regression discontinuity approach, the assumptions that enable identification of causal effects, and methods of implementation. To date, regression discontinuity studies have primarily used linear regression models for continuous outcomes.[2] We show that the design is generalizable to binary, count, and time-to-event outcomes, and to the models that epidemiologists commonly use to analyze them. We then present an application of regression discontinuity to answer a much-debated question: when to start treating HIV patients with antiretroviral therapy (ART).[18] We close by discussing the benefits and limitations of regression discontinuity in comparison with other study designs and suggest some additional applications.

REGRESSION DISCONTINUITY DESIGNS: THEORY AND PRACTICE

When an exposure or treatment is determined by a threshold rule, the regression discontinuity design can be used to estimate causal effects. Threshold rules are common in medicine. Patients are often assigned to a therapeutic regimen if they are identified as "high risk" with respect to a continuous biomarker such as cholesterol, blood glucose, or birth weight.[10] As with most measures in nature, the continuous measures that determine treatment eligibility are subject to random variability due to measurement error, sampling variability, and chance factors that affect biomarkers such as ambient temperature. Random variability implies that patients who score immediately above and below the threshold will be similar, in expectation, on all observed and unobserved pretreatment characteristics, just as in a randomized controlled trial (RCT). Causal effects can be estimated by comparing outcomes in these patients. Threshold rules also appear in nonclinical settings. Eligibility for a program may depend on being born after a certain date,[19] residing in a sufficiently poor county,[14] or on one side of an administrative boundary.[20] Indeed, the assignment variable could be any continuous pretreatment measure including the outcome variable measured at baseline[4] or another measure of risk[7,21]; a baseline covariate that is loosely correlated with the outcome[14,15,19]; or even a random number, in which case regression discontinuity is identical to an RCT.[2] In this paper, we use as a running example the clinical measurement of CD4 counts (cells/μL of blood), which are used to determine eligibility for ART. Measured CD4 counts contain substantial random variability.[22,23] For "true" CD4 counts close to the threshold, these sources of variability will randomly allocate HIV patients to measured CD4 counts above or below the threshold and hence to different probabilities of ART initiation.

Causal Inference in Regression Discontinuity Designs

We provide a brief introduction to regression discontinuity as a method of causal inference, using the potential-outcomes framework.[24] Detailed discussions have been published elsewhere.[2,6,8,25] We assume a binary treatment, although results can be generalized to continuously valued exposures.[14] By definition, causal inference requires comparison of outcomes for the same patients (or other unit of analysis) in 2 states of the world: if treated, $Y_i(1)$, and if not treated, $Y_i(0)$. Only one of these potential outcomes is ever observed: $Y_i=Y_i(1)$ if $T_i=1$ or $Y_i=Y_i(0)$ if $T_i=0$, where $T_i = \{0,1\}$ is the treatment indicator, as assigned. The challenge faced by nonexperimental studies is that if there are unobserved confounders of the relationship between T_i and Y_i, then the potential outcomes will be correlated with treatment assignment and effect estimates will be biased.

Regression discontinuity designs are feasible when the probability of treatment assignment changes discontinuously at some threshold value, c, of a continuous assignment variable, Z_i: $\lim_{Z\downarrow c}\Pr(T_i = 1 \mid Z_i = z) \neq \lim_{Z\uparrow c}\Pr(T_i = 1 \mid Z_i = z)$. If the probability of treatment assignment changes from 0 to 1 at the threshold, then treatment assignment is a deterministic function of Z_i: $T_i = 1[Z_i < c]$, where $1[\cdot]$ is the indicator function; this is known as "sharp regression discontinuity (SRD)." When the probability of treatment changes at the threshold, but not from 0 to 1, this is known as "fuzzy regression discontinuity (FRD)".[5,6]

The key insight that motivates regression discontinuity is that, in a small neighborhood around c, as that range goes toward 0, treatment assignment is ignorable, that is, independent of the potential outcomes, just as in randomized experiments: $\lim_{\varepsilon\to 0} Y_i(0), Y_i(1) \perp T_i \mid c-\varepsilon < Z_i < c+\varepsilon$. This follows from the 2 identifying assumptions of regression discontinuity: first, that Z_i is continuous at c; and second, that the relationship between Z_i and the potential outcomes $Y_i(0), Y_i(1)$ is continuous at c. Under these assumptions, the conditional distribution $f(Y_i(0) \mid Z_i)$ is identical as Z_i approaches c from above and below, and similarly for $f(Y_i(1) \mid Z_i)$. Equivalently, all potential confounders are balanced in a small area around the cutoff. Although continuity at the cutoff may seem like a strong assumption, in fact it follows directly if there is random noise in Z_i (ie, if it is a random variable or if it is measured with error), and patients are unable to manipulate the precise value of Z_i.[2,26] If Z_i is not measured with error (eg, date of birth), if Z_i is noncontinuous (eg, ordinal), or if there is a phase-in region around the cutoff, then regression discontinuity designs can be implemented under a more stringent but often plausible assumption that there are no other reasons for a discontinuity in potential outcomes at the threshold other than treatment assignment.

Most regression discontinuity applications have been concerned with estimating differences in means at the threshold, $E[Y_i(1) \mid Z_i = c] - E[Y_i(0) \mid Z_i = c]$, an average causal effect (ACE). If treatment assignment is deterministic (ie, a "sharp" discontinuity), then patients are assigned to the treatment with certainty if they fall below the threshold and to the control condition if they fall above the threshold: that is, $E[Y_i \mid Z_i] = E[Y_i(1) \mid Z_i]$ when $Z_i < c$, and $E[Y_i \mid Z_i] = E[Y_i(0) \mid Z_i]$ when $Z_i \geq c$. Figure 1 shows the continuous conditional expectation functions for the potential outcomes, $E[Y_i(0) \mid Z_i]$ and $E[Y_i(1) \mid Z_i]$. The solid lines show the observed data, $E[Y_i \mid Z_i]$; the dotted lines show the regions of the potential outcome conditional expectation functions that are not observed. At the threshold, both $E[Y_i(1) \mid Z_i]$ and $E[Y_i(0) \mid Z_i]$ are identified by limits in the observed data. Thus, the sharp regression discontinuity design identifies the average causal effect at the threshold:

$$ACE_{SRD} = lim_{\{z\uparrow c\}} E[Y_i \mid Z_i = z] - lim_{\{z\downarrow c\}} E[Y_i \mid Z_i = z] \qquad (1)$$

Often, treatments are not assigned deterministically but probabilistically (ie, a "fuzzy" discontinuity). This would occur if, for example, clinicians prescribed a therapy to patients based in part on a threshold rule and in part on their clinical judgment. Such is the case with ART for HIV: patients are eligible either if their CD4 count falls below a

threshold value or if they exhibit clinical symptoms that signal the severity of their disease. In the fuzzy regression discontinuity design, Equation 1 is now the intent-to-treat (ITT_{FRD}) effect, that is, the effect of the patient presenting just below the threshold. ITT_{FRD} measures the effect of treatment eligibility, as determined by the threshold rule, and is often of interest in its own right. In particular, ITT_{FRD} can be interpreted as the effect of raising the threshold on outcomes for the full population of patients close to the threshold. In addition, clinicians may be interested in the effect of therapy itself on those induced to take up the treatment because of the threshold rule (so-called compliers). To obtain this complier average causal effect ($CACE_{FRD}$), it is necessary to scale ITT_{FRD} by the difference in the probability of treatment at the cutoff (ie, the Wald instrumental variables estimator, Equation 2). Fuzzy regression discontinuity can be thought of as an instrumental-variables approach, where $\mathbb{1}[Z_i < c \mid Z_i \to c]$ is the instrument.

$$CACE_{RDD} = \frac{lim_{\{z\uparrow c\}} E[Y_i \mid Z_i = z] - lim_{\{z\downarrow c\}} E[Y_i \mid Z_i = z]}{lim_{\{z\uparrow c\}} P(T_i = 1 \mid Z_i = z) - lim_{\{z\downarrow c\}} P(T_i = 1 \mid Z_i = z)} \quad (2)$$

When the denominator of Equation 2 is equal to 1, we are in the sharp regression discontinuity case, and $ITT_{FRD} = CACE_{FRD} = ACE_{SRD}$; when it is 0, there is no discontinuity, and the causal effect is not identified. In our example, $CACE_{FRD}$ measures the casual effect of rapid (vs. deferred) ART initiation only for those induced to initiate because they had an eligible CD4 count; this effect may differ from the (unobserved) treatment effects for patients that would have initiated ART regardless of CD4 count, for example, because of clinical symptoms (so-called always-takers), or patients who would not have initiated ART even if eligible (so-called never-takers).[27] Additionally, identification of $CACE_{FRD}$ requires the assumptions of monotonicity (ie, that no patients who would have taken up ART if ineligible would refuse ART if eligible and vice versa) and of excludability (ie, that $Z_i < c$ may affect Y_i only through T_i). ITT effects have been popular in epidemiology because they do not require these assumptions.[28]

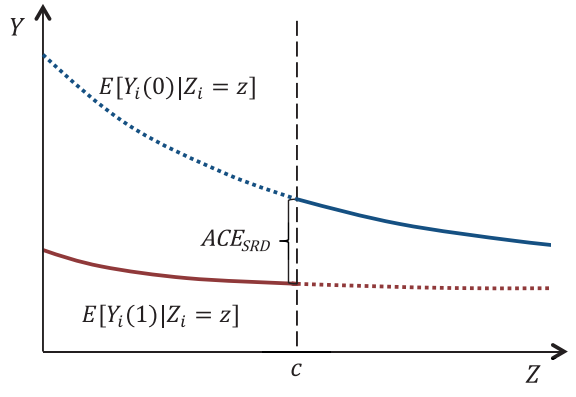

FIGURE 1. Sharp regression discontinuity design. This figure shows the conditional expectation functions for each of the potential outcomes $E[Y_i(1) \mid Z_i = z]$ and $E[Y_i(0) \mid Z_i = z]$. The solid lines show the conditional expectation function of the observed data, $E[Y_i \mid Z_i = z]$.

In both sharp regression discontinuity and fuzzy regression discontinuity designs, causal treatment effects are identified at the threshold. If treatment effects are constant or independent of Z_i, then ITT_{FRD} (and equivalently ACE_{SRD}) is equal to the population average treatment effect identified in an RCT. (In fact, an RCT can be thought of as a discontinuity design in which Z_i is a random number.) If treatment effects are heterogeneous in Z_i (ie, $E[Y_i(0) \mid Z_i]$ and $E[Y_i(1) \mid Z_i]$ are not parallel, as in Figure 1), then the regression discontinuity estimand should be interpreted as a local treatment effect at $Z_i = c$. This local effect is more generalizable than it may first appear. Due to random noise in measurements of Z_i, observations with $Z_i = c$ are drawn from a distribution of true Z_i^*. Thus, treatment effects identified at a single value of the measured Z_i can be thought of as a weighted average across a wider range of true Z_i^*, with the weights proportional to $\Pr(Z_i = c \mid Z_i^* = z)$. Furthermore, even if effects are heterogeneous across the full range of Z_i, they may be approximately constant (on the appropriate scale) for a wide range of values around the threshold; the assumption of constant proportional or additive effects is often invoked in epidemiologic studies (e.g. nonsaturated regression models). The presence of effect heterogeneity close to the threshold can be tested by assessing whether the slope of $E[Y_i \mid Z_i]$ changes at c. We caution, however, that local effects may not be generalizable to populations far from the threshold (eAppendix, http://links.lww.com/EDE/A808). An alternative to local identification at the threshold might be to estimate a global average causal effect by extrapolating the conditional expectation functions across the entire range of Z_i; however, this requires much stronger assumptions to identify causal effects—in particular, that the functional forms of $E[Y_i(1) \mid Z_i]$ and $E[Y_i(0) \mid Z_i]$ are known across the full range of Z_i.[6,21,29] Consistent estimation of the limits in Equations 1 and 2 does not depend on knowledge of the functional form of the conditional expectation functions, so long as one is willing to shrink the bandwidth as the sample size increases.[25]

Estimation in Regression Discontinuity Designs

The task for estimation in regression discontinuity designs is to estimate the limits in Equations 1 and 2: $\lim_{\{z\downarrow c\}} E[Y_i \mid Z_i = z]$ and $\lim_{\{z\uparrow c\}} E[Y_i \mid Z_i = z]$. One approach might be to compare means in a range of Z_i above and below the threshold. However, if the slope of $E[Y_i \mid Z_i]$ is non-zero on either side of the threshold, then these averages will be biased estimates of the true averages at the limit, as $Z_i \to c$. Estimating local linear (or cubic) regression models substantially mitigates this problem.[30] In practice, ACE_{SRD} and ITT_{FRD} estimates are typically formed by fitting parametric functions of $E[Y_i \mid Z_i, Z_i \geq c]$ and $E[Y_i \mid Z_i, Z_i < c]$ for a range

of data around the threshold and taking the difference in the predictions at $Z_i = c$. It is customary to fit models of the form

$$E[Y_i \mid Z_i] = \beta_0 + \beta_1(Z_i - c) + \beta_2 1[Z_i < c] + \beta_3(Z_i - c) * 1[Z_i < c] \quad (3)$$

where β_1 is the slope of the line below the threshold, $\beta_1 + \beta_3$ is the slope of the line above the threshold, and β_2 is the difference at the cutoff.[8] The interaction term allows for the possibility that treatment effects are heterogeneous. Unless the correct functional forms for $E[Y_i(0) \mid Z_i = z]$ and $E[Y_i(1) \mid Z_i = z]$ are known, the finite sample estimate always runs the risk of being biased. However, this problem is considerably reduced by estimating the model using a smaller bandwidth (ie, a narrower window of data $(c-h, c+h)$ around the cutoff) and by assessing the robustness of the results to the inclusion of higher order polynomial terms for Z_i. CACE_{FRD} is estimated by dividing the difference in $E[Y_i \mid Z_i]$ at the threshold by the similarly formed estimate of the difference in $E[T_i \mid Z_i]$ at the threshold.

In regression discontinuity studies, unbiased visual presentation of the data is essential. In particular, the researcher should plot $E[Y_i \mid Z_i]$ and $E[T_i \mid Z_i]$ to show the discontinuity in the outcome and in treatment assignment. Researchers should also provide visual evidence in support of the key identifying assumption (ie, continuity of $f(Y_i(0) \mid Z_i = z)$ and $f(Y_i(1) \mid Z_i = z)$ in Z_i), which results if there is random noise in measurements of Z_i. This assumption has two important implications that can be tested in the data. The first is that the density of the data should be continuous around the threshold; this would be violated if patients (or providers) could precisely manipulate Z_i.[31] The second implication is that baseline covariates should be balanced (ie, continuous) at the threshold. As in RCTs, evidence of balance on baseline observables provides confidence that patients assigned to treatment and control conditions are exchangeable.

Regression Discontinuity with Nonlinear and Censored Regression Models

Regression discontinuity studies have typically used linear regression models, popular among economists.[2,8] There are very few examples of regression discontinuity designs applied to the binary, count, and survival models most often used by epidemiologists.[21,32,33]

The extension of regression discontinuity to nonlinear models is straightforward for ACE_{SRD} and ITT_{FRD}. Continuity in the conditional expectation functions, $E[Y_i(0) \mid Z_i = z]$ and $E[Y_i(1) \mid Z_i = z]$, is sufficient for identification of regression parameters across the class of generalized linear models, which relate the conditional expectation (mean, probability, rate) to a linear model via a continuous link function (such as the log or logit).[34] More generally, continuity in the density functions $f[Y_i(0) \mid Z_i = z]$ and $f[Y_i(1) \mid Z_i = z]$ implies that regression discontinuity can be applied to other estimators that do not rely solely on the mean, such as marginal effects (risk or rate differences) in multiplicative models and quantile regression estimators.[35]

For applications to survival analysis, Equation 3 can be adapted to parametric and semiparametric regression models that specify the hazard, cumulative hazard or survivorship as a function of the assignment variable and time. A common feature of time-to-event data is that some durations are censored, that is, the failure time exceeds the censoring time $Y_i > C_i$. The usual assumption invoked in survival analysis is that the censoring times are noninformative, that is, independent of failure times. For this to hold in regression discontinuity designs, continuity in the distribution of censoring times is required. The inability of agents to manipulate the assignment variable ensures continuity as long as censoring is not a result of treatment assignment. This exclusion is not so innocuous because treatment assignment may influence retention in clinical care and hence the availability of follow-up data. However, this caution applies to longitudinal data collection in general. Validity is enhanced when follow-up data are collected separately from routine monitoring of treated patients.

In fuzzy regression discontinuity designs with nonlinear models, ITT_{FRD} is often of interest and easily estimated. For analysts interested in the effect of the treatment among compliers, rather than the effect of treatment eligibility, CACE_{FRD} can be estimated on the risk difference scale using the simple Wald estimator evaluated at the threshold. This linear estimator is unbiased for nonlinear models without covariates and is identical to the additive structural mean model.[36,37] Complier causal relative risks (CCRR_{FRD}) can be estimated in multiplicative structural mean models.[36–39] Instrumental variables techniques that account for censoring in survival analysis are under development.[40] A simple approach is to use predicted survival probabilities for the numerator in the Wald estimator[40]; under some assumptions, predicted hazards could also be estimated and plugged in (eAppendix, http://links.lww.com/EDE/A808). We note that the null hypothesis $\text{CACE}_{\text{FRD}} = 0$ is equivalent to $\text{ITT}_{\text{FRD}} = 0$ and the variance of CACE_{FRD} is strictly larger than the variance of ITT_{FRD}; if a result is not statistically significant in the ITT framework, it will not be significant after scaling by take-up.

AN APPLICATION OF THE REGRESSION DISCONTINUITY DESIGN: WHEN TO START ANTIRETROVIRAL THERAPY FOR HIV

To illustrate the potential for regression discontinuity in epidemiology, we present a real-life application to a much-debated question: when in the course of HIV disease progression to start life-prolonging ART. We assessed the causal effect of early versus delayed ART eligibility on survival using data from a large cohort of HIV-infected patients in rural South Africa. Our application exploits the threshold rule used to determine ART eligibility during the study period 2007–2011.

Our analysis contributes causal evidence to a question on which experimental evidence is limited. In an RCT

in Haiti, Severe et al[41] found a 75% reduction in mortality among HIV patients who initiated treatment when their CD4 counts were between 200 and 350 cells/μL, rather than waiting for their CD4 counts to fall below 200 cells/μL. Cohen et al[42] found a 41% decrease in clinical events among patients who began treatment between 350 and 550 cells/μL compared with those who delayed therapy until their CD4 count went below 250 cells/μL; however, the study did not have sufficient power to detect differences in survival. No RCT has evaluated the effect of early versus delayed therapy on survival in sub-Saharan Africa where most people receiving ART live. Several large clinical cohort studies have reported higher mortality for patients who initiated ART at lower CD4 counts[43–47]; however, these studies are limited by the potential for bias due to unobserved confounders that determine treatment-seeking behavior and by the exclusion of patients who never initiated ART.

CD4 counts at enrollment in care were obtained for all patients in the Hlabisa HIV Treatment and Care Programme. Dates of ART initiation were obtained for those who initiated therapy.[48] Patients were eligible for ART if their CD4 count was less than 200 cells/μL or if they had stage IV AIDS-defining illness, as per national guidelines.[49] Dates of death were obtained from the Africa Centre for Health and Population Studies, which maintains a demographic surveillance system in the clinical catchment area.[50] Survival data were linked to clinical records by national ID number, full name, age, and sex.[51] The study population included all patients who had a first CD4 count between 1 January 2007 and 11 August 2011—regardless of whether they later initiated ART—and who were under surveillance at that time. Patients with first CD4 counts greater than 350 cells/μL were excluded. Patients were followed from the date of their first CD4 count to their date of death or the date when their vital status was last observed in the population surveillance system. Out of 4391 patients who sought care, 2874 initiated ART and 820 died during 13,139 person-years of follow up. Stata 11 was used for all statistical analysis (StataCorp, College Station, TX).

Figure 2 shows the distribution of baseline CD4 counts among patients in the study sample. Causal inference would be jeopardized if health workers or patients manipulated CD4 counts, for example, in an effort to access treatment earlier. We found no evidence of bunching at the threshold, as would result from manipulation. Further analysis revealed balance in variables observed at baseline (age and sex) at the cutoff (not shown). Figure 3 displays the cumulative probability of ART initiation within 3 and 12 months of a patient's first CD4 count. The probability of rapid ART initiation (within 3 months) was higher for patients presenting below 200 cells/μL; this discontinuity persisted at 1 year.

We first examined the effect of treatment eligibility (CD4 < 200 cells/μL) on mortality in an ITT analysis. The Table presents the results of hazard regression models, with the log-hazard replacing $E[Y_i \mid Z_i]$ in Equation 3. We present estimates

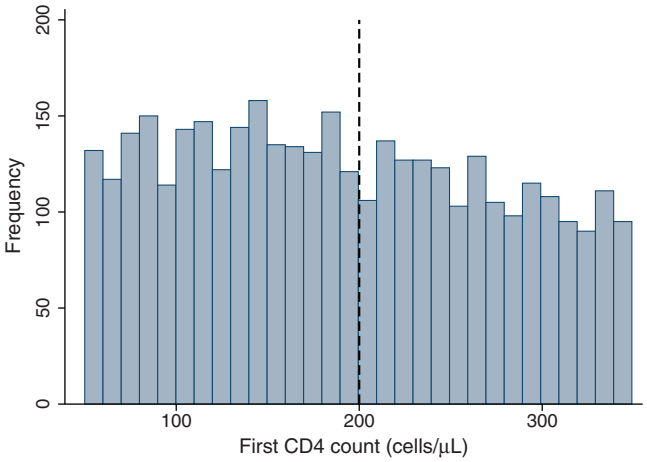

FIGURE 2. Distribution of first CD4 counts in the HIV treatment and care program.

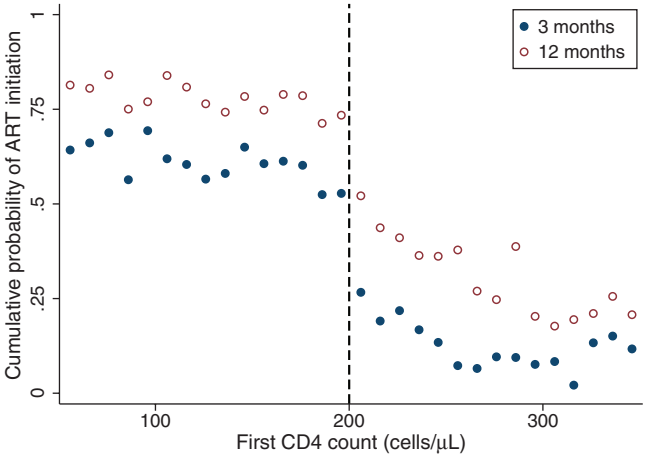

FIGURE 3. First CD4 count and ART initiation. Kaplan–Meier estimates of the probability that a patient initiated ART within 3 and 12 months of first CD4 count in the HIV treatment and care program.

limiting the data to several ranges (bandwidths) around the cutoff. Smaller bandwidths reduce the potential for bias from using a linear function to approximate the relationship between first CD4 count and log-mortality rates; however, this reduction in possible bias is attained at the expense of precision. Figure 4 displays fitted values from model 2a, superimposed over hazards predicted for CD4 count bins of width 10 cells.

In general, mortality was lower for patients presenting with higher initial CD4 counts (Table, Figure 4). However, there was a discontinuity at 200 cells/μL: patients presenting just below the threshold had a 35% lower hazard of death than those presenting just above the threshold (ITT_{FRD} HR = 0.65; model 2a in the Table). This result was robust to varying specifications of the hazard function and statistically significant in models using wider CD4 count bandwidths. In models with smaller bandwidths, the coefficients remained essentially unchanged

TABLE. Intent-to-Treat Estimates: The Causal Effect of ART Eligibility on Mortality[a]

First CD4 Count Range (cells/μL)		Predictor	(a) Exponential HR (95% CI)	(b) Cox HR (95% CI)	Sample No.
1	0–350	D_i	0.59 (0.42–0.83)	0.62 (0.45–0.88)	4,391
		$(Z_i - c)$	0.993 (0.990–0.997)	0.994 (0.990–0.998)	
		$D_i * (Z - c)$	0.996 (0.992–1.000)	0.996 (0.992–1.000)	
2	50–350	D_i	0.65 (0.45–0.94)	0.67 (0.46–0.96)	3,710
		$(Z_i - c)$	0.993 (0.990–0.997)	0.994 (0.990–0.998)	
		$D_i * (Z - c)$	0.997 (0.992–1.001)	0.997 (0.992–1.001)	
3	100–300	D_i	0.66 (0.42–1.04)	0.67 (0.43–1.06)	2,557
		$(Z_i - c)$	0.994 (0.987–1.000)	0.994 (0.988–1.000)	
		$D_i * (Z - c)$	0.997 (0.990–1.005)	0.997 (0.989–1.005)	
4	150–250	D_i	0.68 (0.35–1.32)	0.72 (0.37–1.38)	1,293
		$(Z_i - c)$	0.994 (0.978–1.010)	0.995 (0.979–1.011)	
		$D_i * (Z - c)$	0.997 (0.975–1.020)	0.996 (0.974–1.019)	
5	175–225	D_i	0.54 (0.21–1.41)	0.55 (0.21–1.44)	623
		$(Z_i - c)$	0.988 (0.946–1.032)	0.989 (0.947–1.033)	
		$D_i * (Z - c)$	0.994 (0.932–1.061)	0.992 (0.930–1.059)	

[a] Z_i is first CD4 count. $D_i = 1[Z_i < c]$ is an indicator for whether the patient is eligible for ART according to CD4 count. Table displays the results of 10 regressions—5 ranges of CD4 counts and 2 statistical models for the survival times.

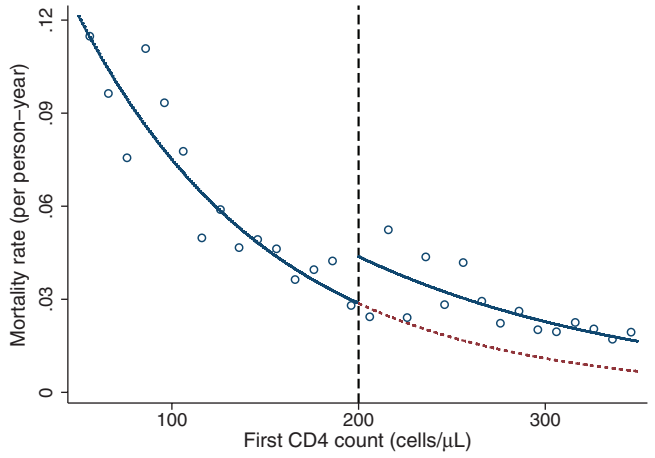

FIGURE 4. First CD4 count and mortality hazard rate. Predicted hazards from the Table, model 2a are displayed as solid lines. Dashed line shows extrapolated prediction if all patients were treatment eligible at first CD4 count. Dots are hazards predicted for CD4 count bins of width 10 cells.

although, as expected, the estimates were less precise. Visual inspection of Figure 4 shows no evidence of misspecification. The hazard ratios on the interaction terms were close to 1.0, suggesting that treatment effects were not heterogeneous close to the threshold (Table).

The ITT_{FRD} estimate is arguably the parameter of interest from a policy perspective: it is the causal effect of ART eligibility for all patients seeking care with CD4 counts close to the threshold. However, clinicians may also be interested in $CACE_{FRD}$, the causal effect of rapid ART initiation on survival among patients who initiated based on their CD4 count. To obtain $CACE_{FRD}$, we scaled the difference in mortality hazards at the threshold by the difference in the probability of ART initiation within 3 months at the threshold. In both cases, we used models with separate linear terms on either side of the threshold, estimated on the range of 50–350 cells. This yielded a causal difference in hazards of $CACE_{FRD} = 0.010 / 0.360 = 0.029$ fewer deaths per person-year for patients who initiated because they were CD4-count eligible, compared with those who were precluded from initiating because they were ineligible. Mortality hazards for treatment and control compliers were calculated to be 0.011 and 0.040, respectively, resulting in a complier causal hazard ratio of 0.28. (See eAppendix for details on these calculations and robustness checks, http://links.lww.com/EDE/A808.) Rapid ART initiation thus causally reduced mortality by 72% among patients who initiated ART because CD4 < 200 cells/μL.

DISCUSSION

Regression discontinuity designs present an opportunity for causal inference in epidemiology when randomization is beyond the control of the researcher. As a quasi-experimental study design, regression discontinuity offers significant benefits over nonexperimental approaches based on regression adjustment or matching. Continuity in the assignment variable at the

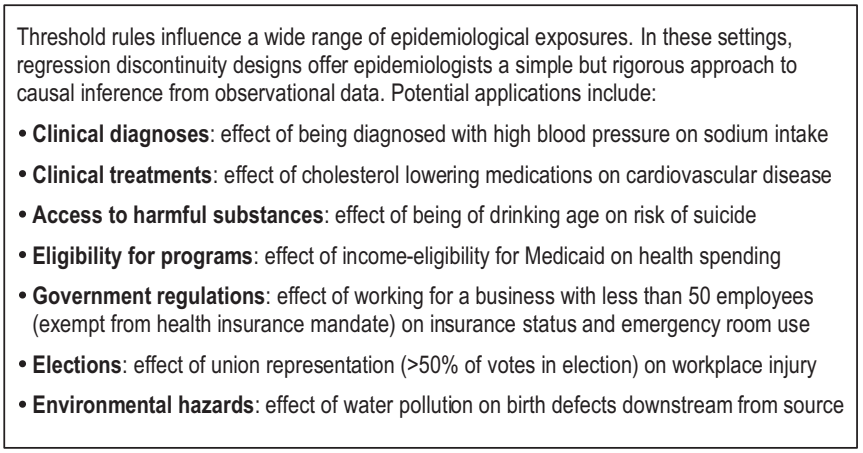

FIGURE 5. Potential applications of regression-discontinuity designs in epidemiology.

threshold breaks all links between treatment assignment and both observed and unobserved confounders. Neither ex-post-covariate adjustment nor assumptions about the absence of residual confounding is required for causal inference, similar to an RCT. Regression discontinuity designs also have important benefits over other quasi-experimental approaches. In most studies using instrumental variables, the assumption that treatment assignment is as-good-as-random is an article of faith; in discontinuity designs, this assumption follows directly from random noise in measurements of the assignment variable and can be assessed through tests of continuity in variables observed at baseline.[2]

There are also important cases where regression discontinuity designs may be preferred to RCTs, such as when it is unethical to deny a randomized intervention to a control group or when an experiment is too expensive or logistically difficult to implement. Additionally, regression discontinuity designs can evaluate the real-world effectiveness of interventions as implemented, providing causal effect estimates that are often more relevant for policy decisions than those derived under the highly controlled conditions of an RCT. Although regression discontinuity designs require larger sample sizes than RCTs to achieve a given level of power,[52] they can often be implemented using routine clinical or administrative data, which are comparatively cheap to collect. Regression discontinuity designs are also more likely to be generalizable to the population seeking care than RCTs with opt-in participant recruitment and a range of participant inclusion and exclusion criteria. Finally, regression discontinuity designs identify a type of causal effect that is of particular interest for policy and clinical practice: the effect for patients near the threshold (which is also the effect of marginally raising or lowering the threshold). In contrast, RCTs estimate average causal effects across a wider range of data and thus do not provide the specific information needed for optimizing treatment thresholds (eAppendix, http://links.lww.com/EDE/A808).

Regression discontinuity designs can be implemented whenever an exposure is assigned—at least in part—by a threshold rule. In spite of many potential applications (Figure 5), regression discontinuity has yet to make substantial inroads in epidemiology.[17] This may be due to (mis)perceptions that the range of applications is limited or that the assumptions required for causal inference are implausible. Some of the early literature on regression discontinuity (and similar designs under other names) proposed that (1) treatment assignment must be based solely on the threshold rule[7,29]; (2) treatment assignment must be under control of the researcher[53]; (3) the functional form of the relationship between the outcome and assignment variable must be known[21,29,54]; (4) treatment effects must be constant[21]; and (5) measurement error in the assignment variable is a source of bias.[54,55] In fact, as described in this paper, assignment need not be deterministic nor under control of the researcher; causal inference can be conducted at the threshold using local linear regression, without functional form assumptions; and treatment effects may be heterogeneous, with the proviso that effects are local to observations near the threshold. Rather than being a threat to validity, random noise in the assignment variable ensures continuity in potential outcomes—the key assumption required for causal inference—and attenuates effect heterogeneity, increasing the generalizability of the estimates.

In our illustration of regression discontinuity, we found large survival benefits to early versus delayed ART initiation at the CD4 count threshold of 200 cells/μL. Our results are similar in magnitude to those reported by Severe et al[41]—the only RCT to report survival impacts of delaying ART until a patient's CD4 count is below 200 cells/μL. Several factors support our interpretation of these results as causal. By design, our analysis is robust to any unobserved factors that are correlated both with timing of treatment initiation and independently correlated with survival. Causal identification depends only on the assumption that these factors are smooth at the threshold, and this is guaranteed by random noise in measurements of CD4 counts. Our results are unlikely to be biased due to systematic misclassification, selection into the sample, or attrition. Mortality data were collected through

semiannual demographic surveillance; CD4 counts were reported directly from the laboratory; and dates of ART initiation were captured from clinical records. The study included all patients who sought care, not just those who initiated ART. And we observed survival in the surveillance system even for patients who were not retained clinically. Although we believe the internal validity of our results to be high, they may not be generalizable to persons who did not seek care and to patients presenting with CD4 counts far from 200 cells/μL.

The beauty of the regression discontinuity design lies in its simplicity: causal effects can be estimated with very few assumptions, and the source of causal identification is transparent and easy to communicate graphically. These qualities stand out compared with other nonexperimental methods that rely on ex-post statistical adjustment. Threshold rules are ubiquitous in clinical practice, in determining eligibility for programs, and exposure to risk factors. Combined with the tremendous growth in new observational data, regression discontinuity designs can play an important role in generating causal evidence on the health effects of interventions and exposures in real-world settings.

ACKNOWLEDGMENTS

Thank you to Joshua Angrist, Matthew Fox, and Guido Imbens, seminar participants at Boston University, University of Witwatersrand, and 2014 International Workshop on HIV/AIDS Observational Databases, and 4 anonymous reviewers for thoughtful feedback on this project; the staff of the Africa Centre for Health and Population Studies and Hlabisa HIV Treatment and Care Programme; and the study participants.

REFERENCES

1. Editorial. Associations are not effects. *Am J Epidemiol*. 1991;133:101–102.
2. Lee DS, Lemieux T. Regression discontinuity designs in economics. *J Econ Lit*. 2010;48:281–355.
3. Thistlewaite D, Campbell D. Regression discontinuity analysis: an alternative to the ex-post facto experiment. *J Educ Psych*. 1960;51: 309–317.
4. Campbell DT, Stanley JC. Experimental and quasi-experimental designs for research on teaching. In: Gage NL, ed. *Handbook of Research on Teaching*. Chicago, IL: Rand McNally & Company; 1963:61–64.
5. Campbell DT. Reforms as experiments. *Am Psychol*. 1969;24:409–429.
6. Trochim W. *Research Design for Program Evaluation: The Regression Discontinuity Design*. Beverly Hills, CA: Sage Publications; 1984.
7. Trochim, W. The Regression Discontinuity Design. In: Sechrest L, Perrin E, and Bunker J, Eds. *Agency for Health Care Policy and Research Conference Proceedings: Research Methodology: Strengthening Causal Interpretations of Nonexperimental Data*. Washington, D.C.: U.S. Department of Health and Human Services; 1990:119–140.
8. Imbens GW, Lemieux T. Regression discontinuity designs: a guide to practice. *J Econom*. 2008;142:615–635.
9. Cook TD. "Waiting for life to arrive": a history of the regression discontinuity design in psychology, statistics and economics. *J Econom*. 2008;1422:636–654.
10. Almond D, Doyle JJ, Kowalski AE, Williams H. Estimating marginal returns to medical care: evidence from at-risk newborns. *Q J Econ*. 2010;125:591–634.
11. Card D, Dobkin C, Maestas N. Does medicare save lives? *Q J Econ*. 2009;124:597–636.
12. Carpenter C, Dobkin C. The effect of alcohol consumption on mortality: regression discontinuity evidence from the minimum drinking age. *Am Econ J Appl Econ*. 2009;1:164–182.
13. Zhao M, Konishi Y, Glewwe P. Does information on health status lead to a healthier lifestyle? Evidence from China on the effect of hypertension diagnosis on food consumption. *J Health Econ*. 2013;32:367–385.
14. Ludwig J, Miller DL. Does Head Start improve children's life chances? Evidence from a regression discontinuity design. *Q J Econ*. 2007; 122:159–208.
15. Snyder SE, Evans WN. The effect of income on mortality: evidence from the Social Security Notch. *Rev Econ Stat*. 2006;88:482–495.
16. Neidell M. Information, avoidance behavior, and health: the effect of ozone on asthma hospitalizations. *J. Hum Resour*. 2009;44:450–478.
17. Moscoe E, Bor J, Bärnighausen T. Regression discontinuity designs in medicine, epidemiology, and public health: a review of current and best practice. *J Clinical Epidemiology*. In press.
18. World Health Organization (WHO). *Antiretroviral Therapy for HIV Infection in Adults and Adolescents: Recommendations for a Public Health Approach: 2013 Revision*. Geneva: WHO; 2013.
19. Bor J. Cash transfers and teen pregnancy in an HIV endemic setting: a regression discontinuity study. In: *Essays on the Economics of HIV/AIDS in Rural South Africa [dissertation]*. Boston: Harvard University, School of Public Health; 2013, 91–150.
20. Chen Y, Ebenstein A, Greenstone M, Li H. Evidence on the impact of sustained exposure to air pollution on life expectancy from China's Huai River policy. *Proc Natl Acad Sci U S A*. 2013;110:12936–12941.
21. Finkelstein MO, Levin B, Robbins H. Clinical and prophylactic trials with assured new treatment for those at greater risk: II. Examples. *Am J Public Health*. 1996;86:696–705.
22. DeGruttola V, Lange N, Dafni U. Modeling the progression of HIV infection. *J Am Stat Assoc*. 1991;86:569–577.
23. Hughes MD, Stein DS, Gundacker HM, Valentine FT, Phair JP, Volberding PA. Within-subject variation in CD4 lymphocyte count in asymptomatic human immunodeficiency virus infection: implications for patient monitoring. *J Infect Dis*. 1994;169:28–36.
24. Rubin D. Estimating causal effects of treatments in randomized and nonrandomized studies. *J Educ Psych*. 1974;66:688–701.
25. Hahn J, Todd P, van der Klaauw W. Identification and estimation of treatment effects with a regression discontinuity design. *Econometrica*. 2001;69:201–209.
26. Lee DS. Randomized experiments from non-random selection in U.S. House elections. *J Econom*. 2008;142:675–697.
27. Angrist JD, Imbens GW, Rubin DB. Identification of causal effects using instrumental variables. *J Am Stat Assoc*. 1996;91:444–455.
28. Robins JM, Tsiatis AA. Correcting for non-compliance in randomized trials using rank preserving structural failure time models. *Commun. Statist. Theory Meth*. 1991;20:2609–2631.
29. Rubin DB. Assignment to treatment group on the basis of a covariate. *J Educ Stat*. 1977;2:1–26.
30. Fan J, Gijbels I. *Local Polynomial Modelling and its Applications*. London: Chapman and Hall; 1996.
31. McCrary J. Manipulation of the running variable in the regression discontinuity design: a density test. *J Econom*. 2008;142:698–714.
32. Berk RA, Rauma D. Capitalizing on nonrandom assignment to treatments: a regression discontinuity evaluation of a crime-control program, *J Am Stat Assoc*. 1983;78:21–27.
33. Berk AR, de Leeuw J. An evaluation of California's inmate classification system using a generalized regression discontinuity design. *J Am Stat Assoc*. 1999;94:1045–1052.
34. Nelder JA, Wedderburn RWM. Generalized linear models. *J Royal Stat Soc Ser A*. 1972;135:370–384.
35. Frandsen BR, Frölich M, Melly B. Quantile treatment effects in the regression discontinuity design. *J Econom*. 2012;168:382–395.
36. Angrist JD. Estimation of limited dependent variable models with dummy endogenous regressors. *J Bus Econ Stat*. 2001;19:2–28.
37. Hernán MA, Robins JM. Instruments for causal inference: an epidemiologist's dream? *Epidemiology*. 2006;17:360–372.
38. Mullahy J. Instrumental-variable estimation of count data models: applications to models of cigarette smoking behavior. *Rev Econ Stat*. 1997;79:586–593.
39. Clarke PS, Windmeijer F. Instrumental variable estimators for binary outcomes. *J Am Stat Assoc*. 2012;107:1638–1652.

40. Abbring JH, van den Berg GJ. Social experiments and instrumental variables with duration outcomes. Tinbergen Institute Discussion Paper 2005-047/3. 2005.
41. Severe P, Juste MA, Ambroise A, et al. Early versus standard antiretroviral therapy for HIV-infected adults in Haiti. *N Engl J Med*. 2010;363:257–265.
42. Cohen MS, Chen YQ, McCauley M, et al.; HPTN 052 Study Team. Prevention of HIV-1 infection with early antiretroviral therapy. *N Engl J Med*. 2011;365:493–505.
43. Palella FJ Jr, Deloria-Knoll M, Chmiel JS, et al.; HIV Outpatient Study Investigators. Survival benefit of initiating antiretroviral therapy in HIV-infected persons in different CD4+ cell strata. *Ann Intern Med*. 2003;138:620–626.
44. Ford N, Kranzer K, Hilderbrand K, et al. Early initiation of antiretroviral therapy and associated reduction in mortality, morbidity and defaulting in a nurse-managed, community cohort in Lesotho. *AIDS*. 2010;24:2645–2650.
45. Kitahata MM, Gange SJ, Abraham AG, et al.; NA-ACCORD Investigators. Effect of early versus deferred antiretroviral therapy for HIV on survival. *N Engl J Med*. 2009;360:1815–1826.
46. Sterne JA, May M, Costagliola D, et al. Timing of initiation of antiretroviral therapy in AIDS-free HIV-1-infected patients: a collaborative analysis of 18 HIV cohort studies. *Lancet*. 2009;373:1352–1363.
47. Fox MP, Sanne IM, Conradie F, et al. Initiating patients on antiretroviral therapy at CD4 cell counts above 200 cells/microl is associated with improved treatment outcomes in South Africa. *AIDS*. 2010;24:2041–2050.
48. Houlihan CF, Bland RM, Mutevedzi PC, et al. Cohort profile: Hlabisa HIV treatment and care programme. *Int J Epidemiol*. 2011;40:318–326.
49. Bor J, Herbst AJ, Newell ML, Bärnighausen T. Increases in adult life expectancy in rural South Africa: valuing the scale-up of HIV treatment. *Science*. 2013;339:961–965.
50. Tanser F, Hosegood V, Bärnighausen T, et al. Cohort profile: Africa Centre Demographic Information System (ACDIS) and population-based HIV survey. *Int J Epidemiol*. 2008;37:956–962.
51. Bor J, Bärnighausen T, Newell C, Tanser F, Newell ML. Social exposure to an antiretroviral treatment programme in rural KwaZulu-Natal. *Trop Med Int Health*. 2011;16:988–994.
52. Goldberger AS. *Selection bias in evaluating treatment effects: some formal illustrations*. *Discussion Paper*. Madison, WI: Institute for Research on Poverty, University of Wisconsin - Madison; 1972.
53. Luft HS. The applicability of the regression discontinuity design in health services research. In: *Agency for Health Care Policy and Research Conference Proceedings: Research Methodology: Strengthening Causal Interpretations of Nonexperimental Data*. Washington, DC: U.S. Department of Health and Human Services; 1990:140–143.
54. Reichardt CS, Trochim W, Cappelleri J. Reports of the death of the regression discontinuity design are greatly exaggerated. *Eval Rev*. 1995;19:39–63.
55. Stanley TD, Robinson A. Sifting statistical significance from the artifact of RD design. *Eval Rev*. 1990;14:166–181.

