Advancing Quality Improvement with Regression Discontinuity Designs

Allan J. Walkey^{1,2,3}, Mari-Lynn Drainoni^{1,2,4,5}, Nicholas Cordella⁶, and Jacob Bor^{7,8}

¹Evans Center for Implementation and Improvement Sciences, ³Section of Pulmonary, Allergy, and Critical Care Medicine, ⁴Section of Infectious Diseases, and ⁶Section of General Internal Medicine, Department of Medicine, Boston University School of Medicine, Boston, Massachusetts; ²Department of Health Law, Policy and Management, ⁷Department of Global Health, and ⁸Department of Epidemiology, Boston University School of Public Health, Boston, Massachusetts; and ⁵VA Center for Healthcare Organization and Implementation Research, Edith Nourse Rogers Memorial VA Hospital, Bedford, Massachusetts

ORCID ID: 0000-0003-4685-6894 (A.J.W.).

Keywords: quality improvement; causal inference; research design

Innovations & Provocations are intended to start conversations. Share your reactions and discuss with other clinicians in the comments section online.

Opportunities to improve healthcare delivery in pulmonary, critical care, allergy, and sleep medicine are legion. Too few patients with chronic obstructive pulmonary disease receive smoking cessation counseling, long-acting bronchodilators, or pulmonary rehabilitation; too many receive inhaled corticosteroids. Few with asthma, chronic obstructive pulmonary disease, or sleep apnea receive appropriate diagnostic testing (1–3). Underuse of low-tidal volume ventilation during acute respiratory distress syndrome is ubiquitous (4, 5). Because of substantial opportunities to improve healthcare delivery, fields of implementation science, improvement science, quality improvement, and others have become increasingly devoted to the study of effective methods for sustainable positive behavior change and process uptake in medicine (6, 7). Although robust implementation of process change is critically important to driving evidence uptake and quality, such rigorous implementation is rarely accompanied by analytical methods that can test

whether changes in process actually led to changes in patient outcomes (8). In short, routine methods to improve healthcare delivery often lack causal inference, and it can be difficult to know whether such efforts are improving patient care.

An Improvement Evaluation “Problem Case”

Unplanned hospital readmissions have long been a target of quality improvement efforts (9, 10). Unsurprisingly, our hospital would like to reduce readmissions. Every department and division within the hospital has instituted a “grassroots” project to reduce readmissions. The hospital administration has also started a large-scale transdepartmental readmission reduction program. The hospital administration readmissions reduction program consists of increasingly resource-intensive interventions implemented according to patient risk for readmission. For example, patients at moderate risk for readmission (as determined by a readmission risk score) receive a follow-up call after hospital discharge; patients at high risk of readmission receive the follow-up call as well as pharmacist-guided medication instructions and subspecialty consults in line with their medical conditions (e.g.,

pulmonology consult for patients admitted with chronic obstructive pulmonary disease [COPD] exacerbation). “Superusers” in the top 5% of risk are also assigned a patient navigator.

The implementation of multiple overlapping programs across different departments seriously complicates analysis of their effectiveness. How can we isolate evaluation of effectiveness of any one readmissions reduction effort (i.e., the hospital administration’s program) in the setting of multiple different simultaneous readmission interventions occurring across multiple hospital departments? How can we tease apart the effectiveness of each of the hospital administration’s interventions? Because readmission reduction efforts require substantial resources, identifying the readmission reduction approach(es) that work is not merely an academic exercise; rather, confidently identifying effective interventions is necessary to efficiently target resources and reduce spending on unsuccessful interventions.

The Traditional “Improvement Evaluation” Armamentarium

Existing implementation and improvement science methods are of little help in evaluating whether our hospital

(Received in original form December 14, 2017; accepted in final form January 10, 2018)

Supported by National Institutes of Health grants K01 HL116768 and R01 HL136660 (A.J.W.).

Correspondence and requests for reprints should be addressed to Allan J. Walkey, M.D., M.Sc., Pulmonary Center, Boston University School of Medicine, 72 East Concord Street, R-304, Boston, MA 02118. E-mail: alwalkey@bu.edu.

Ann Am Thorac Soc Vol 15, No 5, pp 523–529, May 2018

Copyright © 2018 by the American Thoracic Society

DOI: 10.1513/AnnalsATS.201712-942IP

Internet address: www.atsjournals.org

Figure 1. Conceptual frameworks of randomized controlled trials (RCTs) and regression discontinuity designs (RDDs). (A) RCTs allocate interventions on the basis of random assignment. In this example, patients with a predicted readmission risk of more than 20% are randomized to receive an intervention or no intervention. When effective, the intervention shifts the association between predicted and observed mortality rates relative to the counterfactual, unexposed control group. (B) RDD in implementation and improvement science exploits the use of a threshold rule on a continuous assignment variable (in this example, a readmission risk score greater than X) to assign the intervention to patients. The association between the assignment variable (in this example, predicted readmission risk), and the outcome of interest (observed readmission rate) is then evaluated for a discontinuity at the threshold where the intervention was provided. Unlike RCTs, the counterfactual control group is not directly observed, because the intervention is offered to all eligible patients. However, estimates of counterfactual outcomes with and without the intervention are observed immediately above and below the threshold cutoff value, enabling causal inference at the threshold. With additional assumptions, causal effects can be projected to regions beyond the threshold.

administration readmission reduction program actually works. Individual patient-level randomized controlled trials (RCTs) necessitate infrastructure and consent that often shift studies away from the desired real-world context, and they

relegate some patients to “control” groups that do not receive interventions believed to be of benefit. Stepped-wedge or cluster-randomized trials require a large number of implementation sites and complex coordination that is

infeasible for most local attempts to improve healthcare delivery. Simple pre-post designs are confounded by secular trends and unmeasured confounders. Run charts, control charts, and interrupted time-series analyses

Figure 2. Regression discontinuity (RD) designs can be paired with traditional implementation science approaches, such as evaluation of implementation fidelity (the degree to which implementation is delivered as intended), to increase the value of both approaches. For example, when serial assessments of implementation fidelity (e.g., proportion of eligible patients who are successfully contacted by a patient navigator) are paired with RD analyses of implementation effectiveness, the extent of fidelity to the intervention that achieves local effectiveness (e.g., in the figure, 75%) can be determined and guide resource allocation to subsequent/continued implementation efforts.

that account for secular trends require long-term baseline data. Last, all of the aforementioned approaches to measuring change can be biased by postimplementation cointerventions (i.e., other department-level readmission reduction interventions). We need to expand the toolbox for evaluation of quality improvement.

Regression Discontinuity Designs to Achieve Causal Inference in Implementation and Improvement Science

Regression discontinuity designs (RDDs) are an underused methodology in healthcare research that can overcome the limitations of traditional improvement science designs (11). A search of PubMed and the Cochrane Library (performed on May 11, 2017) for the term *regression discontinuity* yielded only four studies that prospectively applied a healthcare quality improvement intervention using RDD

(12–15). We believe that RDDs are a tool that should be used more widely because they help answer a key question: Did a program really change outcomes for patients? This is a problem not just of association but also of so-called causal inference. Lack of attention to causal inference in the evaluation of quality improvement efforts may lead to erroneous conclusions about what works and how to improve.

RDD originated as an approach to evaluating educational interventions. Observational studies in education, as in medicine, suffer from problems of bias from confounding (also called *endogeneity*). The first use of RDD attempted to overcome problems inherent in asking the following research question: Does winning an award result in higher performance, or do the types of people who can win awards just do better, regardless of whether they are recognized? Thistlethwaite and Campbell proposed a novel approach to this problem of confounding in observational studies (16).

They evaluated the effects of a certificate of merit on future career success by leveraging the fact that merit certificates were awarded to all students who scored above a threshold on an aptitude test. The underlying theory was that students scoring just above or just below the award threshold would be similar, effectively randomized by noise in the assessment method used for award selection. Thus, if a “discontinuity” (i.e., change in intercept) existed in the relationship between test scores and the proportion of students reaching career achievements at the award score cutoff, then the quasi-randomization induced by the imperfect assessment for the award could provide causal inference for an effect of the award on future aspirations and achievements. Thistlethwaite and Campbell concluded that receipt of an award early during schooling increased the probability of later receipt of scholarships but did not affect other career plans (16).

At its core, RDD leverages random noise in a continuous variable used to

Figure 3. Example of a plan–do–study–act framework that incorporates regression discontinuity design (RDD) to evaluate both implementation adoption and effectiveness. The boxes outlined in red are steps modified from a traditional plan–do–study–act cycle when RDDs are used to evaluate effectiveness. *Implementation outcomes include adoption, fidelity, cost, penetration, sustainability, evaluated through traditional implementation science methods.

Table 1. Comparison of methods used to evaluate quality improvement programs

Methods	Strengths	Weaknesses
Randomized controlled trial: individual patients	<ul style="list-style-type: none"> • Very strong proof outcome was caused by intervention • Do not need multiple sites • Maximize statistical power for given sample size 	<ul style="list-style-type: none"> • Infrastructure needed for randomization • Ethical concerns for withholding interventions believed to be effective • Consent process shifts away from “real-world” context, owing to nonparticipation • Difficult to assess sustainability • Susceptible to contamination of interventions • Often expensive
Randomized controlled trial: clusters or stepped wedge	<ul style="list-style-type: none"> • Very strong proof outcome was caused by intervention • Site-level randomization decreases contamination 	<ul style="list-style-type: none"> • Large number of sites needed • Difficult sample size/power calculations • Often expensive
Regression discontinuity design	<ul style="list-style-type: none"> • Strong proof that outcome was caused by intervention • Do not need multiple sites • Ability to continuously monitor effectiveness (sustainability) of intervention • Readily integrated with traditional continuous QI strategies (e.g., PDSA) • Can ethically evaluate interventions even if equipoise cannot be claimed, because patients with the highest risk/need/benefit can be targeted • Can be implemented retrospectively using existing clinical thresholds as well as prospectively in QI efforts • Makes explicit the rationing decisions that pervade medical care and offers opportunity to optimize those decisions • Relatively low resource use 	<ul style="list-style-type: none"> • Requires more patients than randomization for same power level • If readily available threshold is unavailable, resources are required to develop and calculate risk scores • May require sophisticated statistical approach • Unclear generalizability away from intervention threshold • Susceptible to contamination of interventions
Interrupted time-series and difference-in-difference designs	<ul style="list-style-type: none"> • Moderate to strong proof that outcome was caused by intervention • Site-level intervention decreases contamination 	<ul style="list-style-type: none"> • Need for longitudinal data collection and/or multiple sites • Sharply demarcated dates separating pre- from postintervention are uncommon in continuous QI • May require sophisticated statistical approaches
Control/run charts	<ul style="list-style-type: none"> • Site-level intervention decreases contamination • Readily integrated into traditional QI strategies (e.g., PDSA) 	<ul style="list-style-type: none"> • Moderate proof that outcome was caused by intervention, owing to unmeasured confounders • Need for longitudinal data collection
Uncontrolled pre–post study	<ul style="list-style-type: none"> • Site-level intervention decreases contamination • Readily integrated into traditional QI strategies (e.g., PDSA) • Relatively low resource use 	<ul style="list-style-type: none"> • Weak proof that outcome was caused by intervention, owing to secular trends and unmeasured confounders

Definition of abbreviations: PDSA = plan–do–study–act; QI = quality improvement.

determine whether patients receive an intervention. Noise inherent in the measurement of the numerical value used to assign someone to an intervention acts similarly to randomization at the threshold used to determine intervention assignment (17–22). In the example of Thistlethwaite and Campbell, students just above and just below the award threshold were effectively randomized to receive the award. Because they were otherwise indistinguishable, these groups of students

near the threshold could be evaluated for the effects of award receipt on future scholastic achievement. Others have used the results of very close elections (e.g., 49.9% vs. 50.1% of the vote) in which the outcome is essentially determined by random noise around the 50% threshold used to determine victory to show strong effects of incumbency on future voting patterns (23, 24). Only a few studies in medicine have applied RDD to clinical settings. A prior study employed cutoffs in

CD4 counts used as treatment thresholds to evaluate the optimal timing of human immunodeficiency virus (HIV) treatment in austere settings. In sub-Saharan Africa, patients with HIV and CD4 counts slightly less than 200 cells per microliter were more than twice as likely to be treated with antiretroviral therapy as those with CD4 counts just above 200 cells per microliter. The observation that patients with a CD4 count just below 200 cells per microliter had a 35% lower risk of death compared

Figure 4. Theoretical results of a regression discontinuity design (RDD) study seeking to evaluate multiple simultaneous interventions to reduce readmissions implemented at different readmission risk cutoffs. In an RDD study, the continuous assignment score (here, risk of readmissions) is plotted on the x-axis against the outcome of interest (here, readmission rate at each level of readmission risk) on the y-axis. A “discontinuity” in the relationship between the score or continuous measure used to assign the intervention and the outcome at the intervention assignment threshold visually demonstrates the effect of the intervention. In this example, there is a reduction of readmissions with interventions 1 and 3, but not intervention 2. Detailed descriptions of statistical approaches to evaluating effects of the intervention in RDD can be found elsewhere (11, 17, 18, 22, 30, 31).

with those with CD4 just above the “treatment threshold” of 200 cells per microliter suggested that earlier initiation of antiretroviral treatment reduced HIV mortality, a finding consistent with RCT results (17). The elegance of RDD as an implementation approach rests in the fact that the process of selecting patients for an intervention using the threshold also imparts the pseudo-randomization that allows for causal inference, as demonstrated

in several studies identifying concordant results between RCT and RDD (25, 26). Figure 1 demonstrates the similar conceptual frameworks underlying RCT and RDD.

Healthcare implementation and improvement sciences offer myriad opportunities to use RDD as a prospective design when assigning “high-risk” patients to new interventions. RDD for quality improvement involves three major steps:

1. Prospectively develop a risk score (27, 28), select a preexisting score, or use a routinely collected continuous clinical variable (e.g., eosinophil cell count) as the basis for the intervention threshold, and choose a threshold value that selects patients to receive the intervention of interest (17);
2. Work with hospital administrators and clinicians to assign the intervention on the basis of the threshold rule; and
3. Collect data for all patients with a risk score calculated and compare patient outcomes just above and below the intervention threshold.

Unlike randomization or stepped-wedge approaches, initial implementation efforts using RDD can be designed to in step 1 to target patients considered to be at higher risk or to have a greater perceived need for the planned intervention. Such an approach may minimize some of the ethical concerns with RCTs in quality improvement studies and avoid the expense and complexity of the stepped-wedge design. Because step 2 of using RDD in quality improvement necessitates evaluation of the success of intervention assignment at the threshold, traditional “implementation outcomes” (e.g., adoption, fidelity) (8) can also be incorporated into an RDD-based approach to quality improvement. For example, RDD introduces novel approaches to assessing the importance of implementation “fidelity,” or the extent to which the intervention is delivered as intended (8). As shown in Figure 2, by pairing assessments of intervention fidelity and effectiveness during repeated intervals over time (e.g., every 3 mo), RDD can be used to identify thresholds of implementation fidelity that are needed to achieve local outcome effectiveness. After evaluation of outcomes in step 3, decisions can be made to expand, contract, maintain, or modify the implementation and/or improvement process, and the evaluation cycle can be repeated. Through these procedures, RDD offers the opportunity to embed rigorous causal inference within implementation–evaluation cycles, enabling data-driven continuous quality improvement. Figure 3 demonstrates how RDD can be integrated within existing strategies for the evaluation of healthcare delivery change, such as “plan–do–study–act” cycles (29). Table 1 compares and contrasts designs and methods for evaluation of quality improvement interventions, including RDD.

Table 2. Examples of questions in pulmonary, critical care, allergy, and sleep quality improvement amenable to study through regression discontinuity designs

Quality Improvement Question	Continuous Measure	Outcomes
Do programs that increase discharge of high-risk ICU patients to intermediate care units reduce ICU readmissions?	ICU readmission risk score	ICU readmission rate
Do programs designed to increase use of anti-IL-5 therapy for patients with uncontrolled asthma and eosinophil counts >500 reduce asthma hospitalizations?	Eosinophil count	Asthma hospitalizations
Do programs designed to increase pulmonary rehabilitation referrals at 6-min walk thresholds reduce COPD exacerbations?	6-min walk test	COPD exacerbations
Does implementation of a rapid response team for patients with high early warning scores reduce cardiac arrests?	Early warning score	Cardiac arrest
Does a program designed to improve positive pressure adherence among patients with high apnea-hypopnea indices improve sleepiness?	Apnea-hypopnea index	Epworth Sleepiness Scale

Definition of abbreviations: COPD = chronic obstructive pulmonary disease; ICU = intensive care unit; IL = interleukin.

Back to the Case: Regression Discontinuity Designs to Evaluate the Effect of a Readmission Reduction Intervention

The hospital administration's selection of different thresholds of readmission risk to implement readmission reduction interventions lends itself to an RDD approach. Despite the presence of multiple simultaneous readmission reduction interventions in other divisions and departments, only the hospital administration interventions were implemented at specific readmission risk score thresholds. Because calculation of the readmission risk is automated within the electronic medical record, clinicians cannot manipulate the risk score to manually "assign" individuals the intervention, a process that would effectively break the quasi-randomization imparted by noise around the selection threshold. Thus, evaluation of readmission rates at each readmission risk threshold allows for independent assessment of the effectiveness of each intervention. Theoretical results of an RDD approach that implements multiple simultaneous readmission reduction interventions at different risk thresholds is shown in Figure 4.

Regression Discontinuity Designs in Pulmonary, Critical Care, Allergy, and Sleep Medicine

Our systematic review identified no RDD designs used to evaluate changes to healthcare delivery within pulmonary, critical care, allergy, or sleep research. Yet, preexisting threshold-based treatment decisions are common in our field and provide opportunities to use RDD (Table 2). For example, the real-world effectiveness of activated protein C could have been assessed with RDD long before the PROWESS SHOCK (Prospective Recombinant Human Activated Protein C

Worldwide Evaluation in Severe Sepsis and Septic Shock) trial showed lack of efficacy (NCT02843685). That is, the Food and Drug Administration–recommended use of activated protein C only in patients with an Acute Physiology and Chronic Health Evaluation II score of 25 or greater, providing a natural "threshold" around which to evaluate effectiveness. Other areas for use of RDD include evaluation of the effectiveness of programs to increase use of anti-interleukin-5 therapy for uncontrolled asthma at eosinophil count thresholds, programs to improve pulmonary rehabilitation referrals among patients with COPD at 6-minute walk thresholds, rapid response teams at early warning score thresholds, early goal-directed therapy at lactate thresholds, or obstructive sleep apnea treatment adherence approaches at apnea–hypopnea index thresholds. Threshold-based treatments emerge nearly anywhere one looks for them. Multiple excellent resources are available that provide detailed guidance on the statistical approaches to RDD (11, 17, 18, 22, 30, 31) in studying these and other problems.

Limitations

RDD is not a panacea. First, statistical power for RDD analysis is relatively low compared with traditional RCTs. Because RDD estimates effects for the area near the threshold cutoff, sample sizes may need to be three- to fourfold greater in RDD than in RCT designs to achieve similar statistical power. RDD may therefore be most useful in larger hospital settings. Second, the balance and continuity of covariates near the assignment cutoff must be assessed for adequacy of quasi-randomization. Third, because effects are identified at the threshold, they are not inherently transportable elsewhere in the distribution. That is, just because something works at the implementation threshold does not mean it will work for patients who have a score far from the threshold. Therefore, sensitivity analysis (and cautious interpretation) is required when seeking to learn about treatment effects away from

the threshold value. Fourth, mandated intervention assignment at a threshold value in a quality improvement framework may meet ethical challenges similar to randomization. However, a nonbinding intervention threshold within a thoughtfully constructed risk score can attenuate ethical concerns regarding withholding of treatments and can maintain inference when analyzed similarly to an instrumental variable. Finally, features other than effectiveness (what RDD examines) are necessary for a full program evaluation, but they are beyond the scope of this paper (8).

Conclusions

Improvement science methods that allow investigators to casually link efforts to change healthcare delivery with patient outcomes are required to increase the "quality" of current implementation and improvement approaches. RDD is an underused approach that results in quasi-randomization through the act of selecting patients for interventions at a specific threshold of a continuous variable, such as a clinical risk score. RDD can achieve causal inference of effectiveness with internal validity that approaches RCT designs, but without the risks of selection bias and external validity threats that accompany the randomization and consent process, without the need for multiple sites required with stepped-wedge or cluster designs, and without requirements of prior secular trends or sharply demarcated intervention dates required of interrupted time-series or control chart analysis. Regression discontinuity is a valuable study design that can be incorporated into various implementation strategies, such as plan–do–study–act cycles, to evaluate implementation efforts that seek to provide interventions to patients who are deemed most in need of a healthcare delivery change. ■

Author disclosures are available with the text of this article at www.atsjournals.org.

References

- Morgenthaler TI, Aronsky AJ, Carden KA, Chervin RD, Thomas SM, Watson NF. Measurement of quality to improve care in sleep medicine. *J Clin Sleep Med* 2015;11:279–291.
- Boulet LP, Bourbeau J, Skomro R, Gupta S. Major care gaps in asthma, sleep and chronic obstructive pulmonary disease: a road map for knowledge translation. *Can Respir J* 2013;20:265–269.
- Bourbeau J, Sebaldt RJ, Day A, Bouchard J, Kaplan A, Hernandez P, et al. Practice patterns in the management of chronic obstructive

- pulmonary disease in primary practice: the CAGE study. *Can Respir J* 2008;15:13–19.
- 4 Bellani G, Laffey JG, Pham T, Fan E, Brochard L, Esteban A, *et al.*; LUNG SAFE Investigators; ESICM Trials Group. Epidemiology, patterns of care, and mortality for patients with acute respiratory distress syndrome in intensive care units in 50 countries. *JAMA* 2016;315:788–800.
- 5 Walkey AJ, Wiener RS. Risk factors for underuse of lung-protective ventilation in acute lung injury. *J Crit Care* 2012;27:323.e1–323.e9.
- 6 Fisher ES, Shortell SM, Savitz LA. Implementation science: a potential catalyst for delivery system reform. *JAMA* 2016;315:339–340.
- 7 Eccles MP, Armstrong D, Baker R, Cleary K, Davies H, Davies S, *et al.* An implementation research agenda. *Implement Sci* 2009;4:18.
- 8 Proctor E, Silmere H, Raghavan R, Hovmand P, Aarons G, Bunker A, *et al.* Outcomes for implementation research: conceptual distinctions, measurement challenges, and research agenda. *Adm Policy Ment Health* 2011;38:65–76.
- 9 Anderson GF, Steinberg EP. Hospital readmissions in the Medicare population. *N Engl J Med* 1984;311:1349–1353.
- 10 Centers for Medicare and Medicaid Services. Readmissions reduction program. 2014 [accessed 2017 Jun 22]. Available from: <https://www.cms.gov/Medicare/Medicare-Fee-for-Service-Payment/AcuteInpatientPPS/Readmissions-Reduction-Program.html>.
- 11 Moscoe E, Bor J, Bärnighausen T. Regression discontinuity designs are underutilized in medicine, epidemiology, and public health: a review of current and best practice. *J Clin Epidemiol* 2015;68:122–133.
- 12 Robinson TE, Zhou L, Kerse N, Scott JD, Christiansen JP, Holland K, *et al.* Evaluation of a New Zealand program to improve transition of care for older high risk adults. *Australas J Ageing* 2015;34:269–274.
- 13 McFarlane WR, Levin B, Travis L, Lucas FL, Lynch S, Verdi M, *et al.* Clinical and functional outcomes after 2 years in the early detection and intervention for the prevention of psychosis multisite effectiveness trial. *Schizophr Bull* 2015;41:30–43.
- 14 Flam-Zalcman R, Mann RE, Stoduto G, Nochajski TH, Rush BR, Koski-Jännes A, *et al.* Evidence from regression-discontinuity analyses for beneficial effects of a criterion-based increase in alcohol treatment. *Int J Methods Psychiatr Res* 2013;22:59–70.
- 15 Yi SW, Shin SA, Lee YJ. Effectiveness of a low-intensity telephone counselling intervention on an untreated metabolic syndrome detected by national population screening in Korea: a non-randomised study using regression discontinuity design. *BMJ Open* 2015;5:e007603.
- 16 Thistlethwaite DL, Campbell DT. Regression-discontinuity analysis: an alternative to the ex post facto experiment. *J Educ Psychol* 1960;51:309–317.
- 17 Bor J, Moscoe E, Mutevedzi P, Newell ML, Bärnighausen T. Regression discontinuity designs in epidemiology: causal inference without randomized trials. *Epidemiology* 2014;25:729–737.
- 18 Bor J, Moscoe E, Bärnighausen T. Three approaches to causal inference in regression discontinuity designs. *Epidemiology* 2015;26:e28–e30.
- 19 Linden A, Adams JL, Roberts N. Evaluating disease management programme effectiveness: an introduction to the regression discontinuity design. *J Eval Clin Pract* 2006;12:124–131.
- 20 Listl S, Jürges H, Watt RG. Causal inference from observational data. *Community Dent Oral Epidemiol* 2016;44:409–415.
- 21 O’Keeffe AG, Geneletti S, Baio G, Sharples LD, Nazareth I, Petersen I. Regression discontinuity designs: an approach to the evaluation of treatment efficacy in primary care using observational data. *BMJ* 2014;349:g5293.
- 22 Oldenburg CE, Moscoe E, Bärnighausen T. Regression discontinuity for causal effect estimation in epidemiology. *Curr Epidemiol Rep* 2016;3:233–241.
- 23 Lee DS, Moretti E, Butler MJ. Do voters affect or elect policies? Evidence from the U.S. House. *Q J Econ* 2004;119:807–859.
- 24 Fowler A, Hall AB. Long-term consequences of election results. *Br J Polit Sci* 2017;47:351–372.
- 25 van Leeuwen N, Lingsma HF, de Craen AJ, Nieboer D, Mooijaart SP, Richard E, *et al.* Regression discontinuity design: simulation and application in two cardiovascular trials with continuous outcomes. *Epidemiology* 2016;27:503–511.
- 26 Cook TD, Shadish WR, Wong VC. Three conditions under which experiments and observational studies produce comparable causal estimates: new findings from within-study comparisons. *J Policy Anal Manage* 2008;27:724–750.
- 27 Sullivan LM, Massaro JM, D’Agostino RB Sr. Presentation of multivariate data for clinical use: the Framingham Study risk score functions. *Stat Med* 2004;23:1631–1660.
- 28 D’Agostino RBS Sr, Vasan RS, Pencina MJ, Wolf PA, Cobain M, Massaro JM, *et al.* General cardiovascular risk profile for use in primary care: the Framingham Heart Study. *Circulation* 2008;117:743–753.
- 29 Langley GL, Moen RD, Nolan KM, Nolan TW, Norman CL, Provost LP. The improvement guide: a practical approach to enhancing organizational performance. 2nd ed. San Francisco: Jossey-Bass; 2009.
- 30 Trochim WMK. Research design for program evaluation: the regression-discontinuity approach. Beverly Hills, CA: Sage Publications; 1984.
- 31 Hahn J, Todd P, Van der Klaauw W. Identification and estimation of treatment effects with a regression-discontinuity design. *Econometrica* 2001;69:201–209.