Natural Experiments: An Overview of Methods, Approaches, and Contributions to Public Health Intervention Research

Peter Craig, Srinivasa Vittal Katikireddi, Alastair Leyland, and Frank Popham

MRC/CSO Social and Public Health Sciences Unit, University of Glasgow, Glasgow G2 3QB, United Kingdom; email: peter.craig@glasgow.ac.uk, vittal.katikireddi@glasgow.ac.uk, alastair.leyland@glasgow.ac.uk, frank.popham@glasgow.ac.uk

Annu. Rev. Public Health 2017. 38:39–56

First published online as a Review in Advance on January 11, 2017

The *Annual Review of Public Health* is online at publhealth.annualreviews.org

https://doi.org/10.1146/annurev-publhealth-031816-044327

Keywords

population health interventions, evaluation methods, causal inference

Abstract

Population health interventions are essential to reduce health inequalities and tackle other public health priorities, but they are not always amenable to experimental manipulation. Natural experiment (NE) approaches are attracting growing interest as a way of providing evidence in such circumstances. One key challenge in evaluating NEs is selective exposure to the intervention. Studies should be based on a clear theoretical understanding of the processes that determine exposure. Even if the observed effects are large and rapidly follow implementation, confidence in attributing these effects to the intervention can be improved by carefully considering alternative explanations. Causal inference can be strengthened by including additional design features alongside the principal method of effect estimation. NE studies often rely on existing (including routinely collected) data. Investment in such data sources and the infrastructure for linking exposure and outcome data is essential if the potential for such studies to inform decision making is to be realized.

INTRODUCTION

Natural experiments (NEs) have a long history in public health research, stretching back to John Snow's classic study of London's cholera epidemics in the mid-nineteenth century. Since the 1950s, when the first clinical trials were conducted, investigators have emphasized randomized controlled trials (RCTs) as the preferred way to evaluate health interventions. Recently, NEs and other alternatives to RCTs have attracted interest because they are seen as the key to evaluating large-scale population health interventions that are not amenable to experimental manipulation but are essential to reducing health inequalities and tackling emerging health problems such as the obesity epidemic (15, 27, 40, 68, 76).

We follow the UK Medical Research Council guidance in defining NEs broadly to include any event not under the control of a researcher that divides a population into exposed and unexposed groups (16). NE studies use this naturally occurring variation in exposure to identify the impact of the event on some outcome of interest. Our focus here is on public health and other policy interventions that seek to improve population health or which may have important health impacts as a by-product of other policy goals. One key evaluation challenge is selective exposure to the intervention, leading exposed individuals or groups to differ from unexposed individuals or groups in characteristics associated with better or worse outcomes. Understanding and modeling the process(es) determining exposure to the intervention are therefore central to the design and conduct of NE studies.

Some authors define NEs more narrowly to include only those in which the process that determines exposure (often referred to as the assignment or data-generating process) is random or as-if random (22, pp. 15–16). Truly random assignment, although not unknown (14), is extremely rare in policy and practice settings. As-if randomness lacks a precise definition, and the methods proposed to identify as-if random processes (such as a good understanding of the assignment process and checks on the balance of covariates between exposed and unexposed groups) are those used to assess threats to validity in any study that attempts to make causal inferences from observational data. In the absence of a clear dividing line, we prefer to adopt a more inclusive definition and to assess the plausibility of causal inference on a case-by-case basis.

In the next section, we set out a general framework for making causal inferences in experimental and observational studies. The following section discusses the main approaches used NE studies to estimate the impact of public health interventions and to address threats to the validity of causal inferences. We conclude with brief proposals for improving the future use of NEs.

CAUSAL INFERENCE IN TRIALS AND OBSERVATIONAL STUDIES

The potential outcomes model provides a useful framework for clarifying similarities and differences between true experiments on the one hand and observational studies (including NEs) on the other hand (51). Potential outcomes refer to the outcomes that would occur if a person (or some other unit) were exposed simultaneously to an intervention and a control condition. As only one of those outcomes can be observed, causal effects must be inferred from a comparison of average outcomes among units assigned to an intervention or to a control group. If assignment is random, the groups are said to be exchangeable and the intervention's average causal effect can be estimated from the difference in the average outcomes for the two groups. In a well-conducted RCT, randomization ensures exchangeability. In an observational study, knowledge of the assignment mechanism can be used to make the groups conditionally exchangeable, for example, by controlling for variables that influence both assignment and outcomes to the extent that these variables are known and accurately measured (34).

As well as showing why a control group is needed, this framework indicates why an understanding of the assignment process is so important to the design of an NE study. The methods discussed in the next section can be seen as different ways of achieving conditional exchangeability. The framework also usefully highlights the need to be clear about the kind of causal effect being estimated and, in particular, whether it applies to the whole population (such as an increase in alcohol excise duty) or a particular subset (such as a change in the minimum legal age for purchasing alcohol). A comparison of outcomes between groups assigned to the intervention or control condition provides an estimate of the effect of assignment, known as the intention-to-treat (ITT) effect, rather than the effect of the intervention itself. The two are necessarily the same only if there is perfect compliance. Some methods, such as fuzzy regression discontinuity (RD) and instrumental variables (IVs), estimate a different effect, the complier average causal effect (CACE), which is the effect of the intervention on those who comply with their allocation into the control or intervention group (11). Under certain assumptions, the CACE is equivalent to the ITT effect divided by the proportion of compliers. Which effect is relevant will depend on the substantive questions the study is asking. If the effect of interest is the ITT effect, as in a pragmatic effectiveness trial or a policy evaluation in which decision makers wish to know about the effect across the whole population, methods that estimate a more restricted effect may be less useful (20).

A related issue concerns extrapolation—using results derived from one population to draw conclusions about another. In a trial, all units have a known probability of being assigned to the intervention or control group. In an observational study, where exchangeability may be achieved by a method such as matching or by conditioning on covariates, intervention groups may be created whose members in practice have no chance of receiving the treatment (55). The meaning of treatment effects estimated in this way is unclear. Extrapolation may also be a problem for some NE studies, such as those using RD designs, which estimate treatment effects at a particular value of a variable used to determine assignment. Effects of this kind, known as local average treatment effects, may be relevant to the substantive concerns of the study, but researchers should bear in mind how widely results can be extrapolated, given the nature of the effects being estimated.

Table 1 summarizes similarities and contrasts between RCTs, NEs, and nonexperimental observational studies.

METHODS FOR EVALUATING NATURAL EXPERIMENTS

Key considerations when choosing an NE evaluation method are the source of variation in exposure and the size and nature of the expected effects. The source of variation in exposure may be quite simple, such as an implementation date, or quite subtle, such as a score on an eligibility test. Interventions that are introduced abruptly, that affect large populations, and that are implemented where it is difficult for individuals to manipulate their treatment status are more straightforward to evaluate. Likewise, effects that are large and follow rapidly after implementation are more readily detectable than more subtle or delayed effects. One example of the former is a study that assessed the impact of a complete ban in 1995 on the import of pesticides commonly used in suicide in Sri Lanka (32). Suicide rates had risen rapidly since the mid-1970s, then leveled off following a partial ban on pesticide imports in the early 1980s. After the complete ban, rates of suicide by self-poisoning fell by 50%. The decrease was specific to Sri Lanka, was barely offset by an increase in suicide by other methods, and could not be explained by changes in death recording or by wider socioeconomic or political trends.

Although NE studies are not restricted to interventions with rapid, large effects, more complicated research designs may be needed where effects are smaller or more gradual. **Table 2** summarizes approaches to evaluating NEs. It includes both well-established and widely used

Table 1 Similarities and differences between RCTs, NEs, and observational studies

Type of study	Is the intervention well defined?	How is the intervention assigned?	Does the design eliminate confounding?	Do all units have a nonzero chance of receiving the treatment?
RCTs	A well-designed trial should have a clearly defined intervention described in the study protocol.	Assignment is under the control of the research team; units are randomly allocated to intervention and control groups.	Randomization means that, in expectation, there is no confounding, but imbalances in covariates could arise by chance.	Randomization means that every unit has a known chance of receiving the treatment or control condition.
NEs	Natural experiments are defined by a clearly identified intervention, although details of compliance, dose received, etc., may be unclear.	Assignment is not under the control of the research team; knowledge of the assignment process enables confounding due to selective exposure to be addressed.	Confounding is likely due to selective exposure to the intervention and must be addressed by a combination of design and analysis.	Possibility of exposure may be unclear and should be checked. For example, RD designs rely on extrapolation but assume that at the discontinuity units could receive either treatment or no treatment.
Nonexperimental observational studies	There is usually no clearly defined intervention, but there may be a hypothetical intervention underlying the comparison of exposure levels.	There is usually no clearly defined intervention and there may be the potential for reverse causation (i.e., the health outcome may be a cause of the exposure being studied) as well as confounding.	Confounding is likely due to common causes of exposure and outcomes and can be addressed, in part, by statistical adjustment; residual confounding is likely, however.	Possibility of exposure is rarely considered in observational studies so there is a risk of extrapolation unless explicitly addressed.

Abbreviations: NE, natural experiment; RCT, randomized controlled trial; RD, regression discontinuity.

methods such as difference-in-differences (DiD) and interrupted time series (ITS), as well as more novel approaches such as synthetic controls. Below, we describe these methods in turn, drawing attention to their strengths and limitations and providing examples of their use.

Regression Adjustment

Standard multivariable models, which control for observed differences between intervention and control groups, can be used to evaluate NEs when no important differences in unmeasured characteristics between intervention and control groups are expected (see Model 1 in Appendix 1). Goodman et al. used data from the UK Millennium Cohort Study to evaluate the impact of a school-based cycle training scheme on children's cycling behavior (29). The timing of survey fieldwork meant that some interviews took place before and others after the children received training. Poisson models were used to estimate the effect of training on cycling behaviors, with adjustment for a wide range of potential confounders. Previous evaluations that compared children from participating and nonparticipating schools found substantial effects on cycling behavior. In contrast, this study found no difference, suggesting that the earlier findings reflected the selective

Table 2 Approaches to evaluating NEs

Description	Advantages/disadvantages	Examples
Prepost		
Outcomes of interest compared in a population pre- and postexposure to the intervention	Requires data in only a single population whose members serve as their own controls Assumes that outcomes change only as a result of exposure to the intervention	Effect of pesticide import bans and suicide in Sri Lanka (32)
Regression adjustment		
Outcomes compared in exposed and unexposed units, and a statistical model fitted to take account of differences between the groups in characteristics thought to be associated with variation in outcomes	Takes account of factors that may cause both the exposure and the outcome Assumes that all such factors have been measured accurately so that there are no unmeasured confounders	Effect of repeal of handgun laws on firearm-related murders in Missouri (17) Effect of a cycle training scheme on cycling rates in British schoolchildren (29)
Propensity scores		
Likelihood of exposure to the intervention calculated from a regression model and either used to match exposed and unexposed units or fitted in a model to predict the outcome of interest	Allows balanced comparisons when many factors are associated with exposure Assumes that all such factors have been measured accurately so that there are no unmeasured confounders	Effect of the Sure Start scheme in England on the health and well-being of young children (54)
Difference-in-differences		
Change in the outcome of interest pre- and postintervention compared in exposed and unexposed groups	Uses differencing procedure to control for variation in both observed and unobserved fixed characteristics Assumes that there are no group-specific trends that may influence outcomes—the parallel trends assumption	Effect of traffic policing on road traffic accidents in Oregon (19) Effect of paid maternity leave on infant mortality in LMICs (58)
Interrupted time series		
Trend in the outcome of interest compared pre- and postintervention, using a model that accounts for serial correlation in the data and can identify changes associated with introduction of the intervention. Change also compared in exposed and unexposed populations in controlled time series analyses	Provides a powerful and flexible method for dealing with trend data Requires substantial numbers of pre- and postintervention data points; controlled time series analyses may not be possible if the trends in the intervention and control area differ markedly	Effect of a multibuy discount ban on alcohol sales in Scotland (64) Effect of 20-mph zones on road traffic casualties in London, UK (31)
Synthetic controls		
Trend in the outcome of interest compared in an intervention area and a synthetic control area, representing a weighted composite of real areas that mimics the preintervention trend	Does not rely on the parallel trends assumption or require identification of a closely matched geographical control May not be possible to derive a synthetic control if the intervention area is an outlier	Effect of a ban on the use of *trans*-fats on heart disease in Denmark (62) Effect of antitobacco laws on tobacco consumption in California (1)

(*Continued*)

Table 2 (*Continued*)

Description	Advantages/disadvantages	Examples
Regression discontinuity		
Outcomes compared in units defined by scores just above and below a cutoff in a continuous forcing variable that determines exposure to an intervention	Units with scores close to the cutoff should be very similar to one another, especially if there is random error in the assignment variable; some key assumptions can be tested directly Estimates the effects for units with scores close to the cutoff, which may not be generalizable to units with much higher or lower scores on the forcing variable; there is a trade-off between statistical power (which requires including as many people as possible near the cutoff) and minimizing potential confounding (by including only those very close to the cutoff)	Effect of the Head Start program on child mortality in the United States (52) Effects of conditional cash transfers on rates of overweight/obesity in Mexico (6)
Instrumental variables		
A variable associated with exposure to the intervention, but not with other factors associated with the outcome of interest, used to model the effect of the intervention	An instrumental variable that satisfies these assumptions should provide an unconfounded estimate of the effect of the intervention Such variables are rare, and not all of the assumptions can be tested directly	Effect of food stamps on food insecurity (78) Effect of community salons on social participation and self-rated health among older people in Japan (39)

Abbreviations: LMIC, low- and middle-income countries; NE, natural experiment.

provision of training. The key strength of the study by Goodman et al. is the way the timing of data gathering in relation to exposure created well-balanced intervention and control groups. Without this overlap between data gathering and exposure to the intervention, there was a significant risk that unobserved differences between the groups would bias the estimates, despite adjusting for a wide range of observed confounders.

Propensity Score–Based Methods

In a well-conducted RCT, random allocation ensures that intervention and control arms are balanced in terms of both measured and unmeasured covariates. In the absence of random allocation, the propensity score attempts to recreate the allocation mechanism, defined as the conditional probability of an individual being in the intervention group, given a number of covariates (65).

The propensity score is typically estimated using logistic regression, based on a large number of covariates, although alternative estimation methods are available. There are four principal ways to use the propensity score to obtain an estimated treatment effect: matching, stratification, inverse probability weighting, and covariate adjustment (7). Each method will adjust for differences in characteristics of the intervention and control groups and, in so doing, minimize the effects of confounding. The propensity score, however, is constrained by the covariates available and the extent to which they can collectively mimic the allocation to intervention and control groups.

Understanding the mechanism underlying allocation to intervention and control groups is key when deriving the propensity score. Sure Start Local Programmes (SSLPs), area-based interventions designed to improve the health and well-being of young children in England, were an

example where, on an ITT basis, exposure to the intervention was determined by area of residence and would apply to everyone living in the area regardless of individual characteristics. Melhuish et al. (54) therefore constructed a propensity score at the area level, based on 85 variables, to account for differences between areas with and without SSLPs. Analysis was undertaken on individuals clustered within areas, stratified by the propensity of an area to receive the SSLP. The most deprived areas were excluded from the analysis because there were insufficient comparison areas.

Advantages of using the propensity score over simple regression adjustment include the complexity of the propensity score that can be created (through, for example, including higher-order terms and interactions), the ease of checking the adequacy of the propensity score as opposed to checking the adequacy of a regression model, and the ability to examine the extent to which intervention and control groups overlap in key covariates (7, 18), and thereby avoid extrapolation. Although in statistical terms the use of propensity scores may produce results that differ little from those obtained through traditional regression adjustment (70), they encourage clearer thinking about study design and particularly the assignment mechanism (66). When membership of the treatment and control groups varies over time, inverse probability weighting can be used to account for time-varying confounding (34), as in the study by Pega et al. (59) of the cumulative impact of tax credits on self-rated health.

Difference-in-Differences

In its simplest form, the DiD approach compares change in an outcome among people who are newly exposed to an intervention with change among those who remain unexposed. Although these differences could be calculated from a 2×2 table of outcomes for each group at each time point, the effect is more usefully estimated from a regression with terms for group, period, and group-by-period interaction. The coefficient of the interaction term is the DiD estimator (Model 2 in Appendix 1).

DiD's strength is that it controls for unobserved as well as observed differences in the fixed (i.e., time-invariant) characteristics of the groups and is therefore less prone to omitted variable bias caused by unmeasured confounders or measurement error. The method relies on the assumption that, in the absence of the intervention, preimplementation trends would continue. This common trends assumption may be violated by differential changes in the composition of the intervention or control groups or by other events (such as the introduction of another intervention) that affect one group but not the other. With data for multiple preimplementation time points, the common trends assumption can be investigated directly, and it can be relaxed by extending the model to include terms for group-specific trends. With more groups and time points, the risk that other factors may influence outcomes increases, but additional terms can be included to take account of time-varying characteristics of the groups.

De Angelo & Hansen (19) used a DiD approach to estimate the effectiveness of traffic policing in reducing road traffic injuries and fatalities by taking advantage of a NE provided by the state of Oregon's failure to agree on a budget in 2003, which led to the layoff of more than one-third of Oregon's traffic police force. A comparison of injury and fatality rates in Oregon with rates in two neighboring states before and after the layoff indicated that, after allowing for other factors associated with road traffic incidents, such as the weather and the number of young drivers, less policing led to a 12–14% increase in fatalities. Whereas De Angelo & Hansen's study focused on an intervention in a single area, Nandi and colleagues (58) applied DiD methods to estimate the impact of paid maternity leave across a sample of 20 low- and middle-income countries.

DiD methods are not limited to area-based interventions. Dusheiko et al. (23) used the withdrawal of a financial incentive scheme for family doctors in the English National Health Service to

identify whether it led to treatment rationing. Recent developments, such as the use of propensity scores, rather than traditional covariate adjustment, to account for group-specific time-varying characteristics, add additional complexity, but combining DiD with other approaches in this way may further strengthen causal inference.

Interrupted Time Series

Alongside DiD, ITS methods are among the most widely applied approaches to evaluating NEs. An ITS consists of a sequence of count or continuous data at evenly spaced intervals over time, with one or more well-defined change points that correspond to the introduction of an intervention (69). There are many approaches to analyzing time series data (44). A straightforward approach is to use a segmented regression model, which provides an estimate of changes in the level and trend of the outcome associated with the intervention, controlling for preintervention level and trend (43, 75). Such models can be estimated by fitting a linear regression model, including a continuous variable for time since the start of the observation period, a dummy variable for time period (i.e., before/after intervention), and a continuous variable for time postintervention (Model 3 in Appendix 1). The coefficients of these variables measure the preintervention trend, the change in the level of the outcome immediately postintervention, and the change in the trend postintervention. Additional variables can be added to identify the effects of interventions introduced at other time points or to control for changes in level or trend of the outcome due to other factors. Lags in the effect of the intervention can be accounted for by omitting outcome values that occur during the lag period or by modeling the lag period as a separate segment (75). Successive observations in a time series are often related to one another, a problem known as serial autocorrelation. Unless autocorrelation is addressed, the standard errors will be underestimated, but models that allow for autocorrelation can be fitted using standard statistical packages.

By accounting for preintervention trends, well-conducted ITS studies permit stronger causal inference than do cross-sectional or simple prepost designs, but they may be subject to confounding by cointerventions or changes in population composition. Controlled ITS designs, which compare trends in exposed and unexposed groups or in outcomes that are not expected to change as a result of the intervention, can be used to strengthen causal inference still further; in addition, standardization can be used to control for changes in population composition. A common shortcoming in ITS analyses is a lack of statistical power (61). Researchers have published a range of recommendations for the number of data points required, but statistical power also depends on the expected effect size and the degree of autocorrelation. Studies with few data points will be underpowered unless the effect size is large. Zhang et al. (79) and Mcleod & Vingilis (53) provide methods for calculating statistical power for ITS studies.

Robinson et al. (64) applied controlled ITS methods to commercially available alcohol sales data to estimate the impact of a ban on the offer of multipurchase discounts by retailers in Scotland. Because alcohol sales vary seasonally, the researchers fitted models that took account of seasonal autocorrelation, as well as trends in sales in England and Wales where the legislation did not apply. After adjusting for sales in England and Wales, the study found a 2% decrease in overall sales, compared with a previous study's finding of no impact using DiD methods applied to self-reported alcohol purchase data.

Synthetic Controls

The difficulty of finding control areas that closely match the background trends and characteristics of the intervention area is a significant challenge in many NE studies. One solution is to use a synthetic combination of areas rather than the areas themselves as controls. Methods for deriving

synthetic controls and using them to estimate the impact of state-, region-, or national-level policies were developed by political scientists (1–4) and are now being applied to many health and social policies (8, 9, 17, 30, 45, 62, 67).

A synthetic control is a weighted average of control areas that provides the best visual and statistical match to the intervention area on the preintervention values of the outcome variable and of predictors of the outcome. Although the weights are based on observed characteristics, matching on the outcome in the preintervention period minimizes differences in unobserved fixed and time-varying characteristics. The difference between the postintervention trend in the intervention and synthetic control provides the effect estimate. Software to implement the method is available in a number of statistical packages (2).

Abadie et al. (1) used synthetic controls to evaluate a tobacco control program introduced in California in 1988, which increased tobacco taxes and earmarked the revenues for other tobacco control measures. The comparator was derived from a donor pool of other US states, excluding any states that had implemented extensive tobacco control interventions. A weighted combination of five states, based on pre-1988 trends in cigarette consumption and potential confounders, formed the synthetic control. Comparison of the postintervention trends in the real and synthetic California suggested a marked reduction in tobacco consumption as a result of the program.

The synthetic control method can be seen as an extension of the DiD method, with a number of advantages. In particular, it relaxes the requirement for a geographical control that satisfies the parallel trends assumption and relies less on subjective choices of control areas. A practical limitation, albeit one that prevents extrapolation, is that if the intervention area is an outlier, for example if California's smoking rate in 1988 was higher than those of all other US states, then no combination of areas in the donor pool can provide an adequate match. Another limitation is that conventional methods of statistical inference cannot be applied, although Abadie et al. (1) suggest an alternative that compares the estimated effect for the intervention area with the distribution of placebo effects derived by comparing each area in the donor pool with its own synthetic control.

Instrumental Variables

IV methods address selective exposure to an intervention by replacing a confounded direct measure of exposure with an unconfounded proxy measure, akin to treatment assignment in an RCT (33). To work in this way, an IV must be associated with exposure to the intervention, must have no association with any other factors associated with exposure, and must be associated with outcomes only through its association with exposure to the intervention (**Figure 1**).

IVs that satisfy the three conditions offer a potentially valuable solution to the problem of unobserved as well as observed confounders. Estimating an intervention's effect using IVs can be viewed as a two-stage process (Models 5.1 and 5.2 in Appendix 1). In the first stage, a prediction of treatment assignment is obtained from a regression of the treatment variable on the instruments. Fitted values from this model replace the treatment variable in the outcome regression (41).

IVs are widely used in econometric program evaluation and have attracted much recent interest in epidemiology, particularly in the context of Mendelian randomization studies, which use genetic variants as instruments for environmental exposures (25, 36, 48). IV methods have not yet been widely used to evaluate public health interventions because it can be difficult to find suitable instruments and to demonstrate convincingly, using theory or data, that they meet the second and third conditions above (35, 71). A recent example is the study by Ichida et al. (39) of the

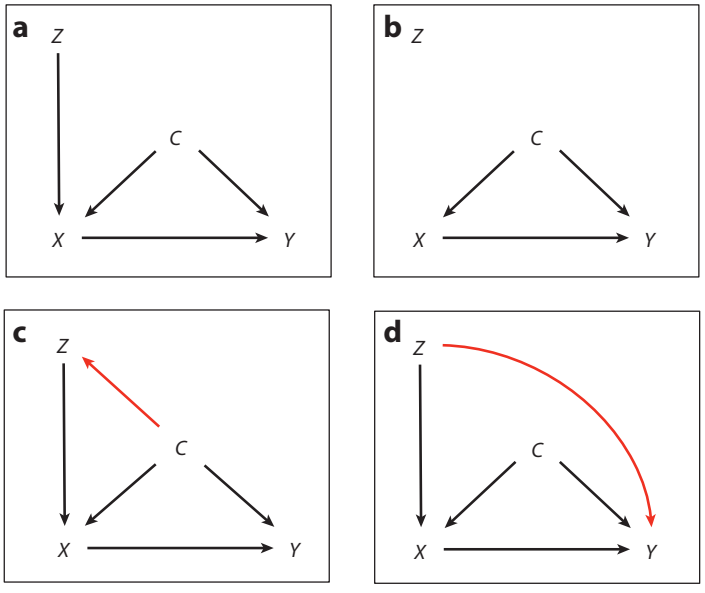

Figure 1

Directed acyclic graphs illustrating the assumptions of instrumental variable (IV) analysis. (*a*) The variable *Z* is associated with outcome *Y* only through its association with exposure *X*, so it can be considered a valid instrument of *X*. (*b*) *Z* is not a valid instrument owing to a lack of any association with outcome *Y*. (*c*) *Z* is not a valid instrument owing to its association with confounder *C*. (*d*) *Z* is not a valid instrument owing to its direct association with *Y*.

effect of community centers on improving social participation among older people in Japan, using distance to the nearest center as an instrument for intervention receipt. Another study, by Yen et al. (78), considers the effect of food stamps on food insecurity, using a range of instruments, including aspects of program administration that might encourage or discourage participation in the food stamp program. Given the potential value of IVs, as one of a limited range of approaches for mitigating the problems associated with unobserved confounders, and their widespread use in related fields, they should be kept in mind should opportunities arise (35).

Regression Discontinuity

Age, income, and other continuous variables are often used to determine entitlement to social programs, such as means-tested welfare benefits. The RD design uses such assignment rules to estimate program impacts. RD is based on the insight that units with values of the assignment variable just above or below the cutoff for entitlement will be similar in other respects, especially if there is random error in the assignment variable (11). This similarity allows the effect of the program to be estimated from a regression of the outcome on the assignment variable (often referred to as the running or forcing variable) and a dummy variable denoting exposure (treatment), with the coefficient of the dummy identifying the treatment effect (Model 4 in Appendix 1). Additional terms are usually included in the model to allow slopes to vary above and below the cutoff, allow for nonlinearities in the relationship between the assignment and outcome variables, and deal with residual confounding.

Visual checks play an important role in RD studies. Plots of treatment probability (**Figure 2**) and outcomes against the assignment variable can be used to identify discontinuities that indicate a treatment effect, and a histogram of the assignment variable can be plotted to identify bunching

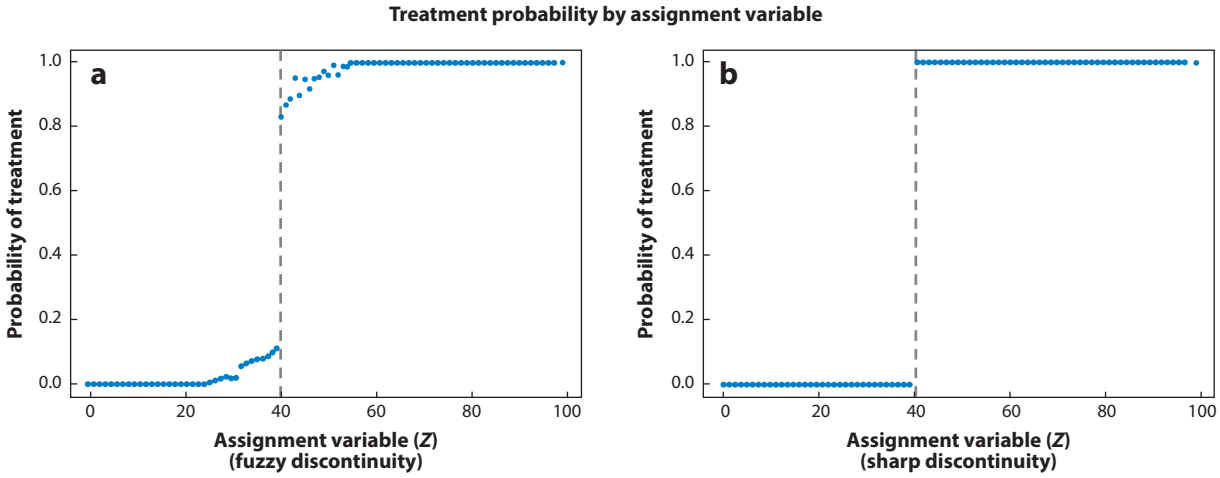

Figure 2

Probability of receiving treatment in fuzzy and sharp regression discontinuity designs. (*a*) A fuzzy regression discontinuity: probability of treatment changes gradually at values of the assignment variable close to the cutoff. (*b*) A sharp regression discontinuity: probability of treatment changes from 0 to 1 at the cutoff. Source: Reproduced from Moscoe (2015) (57) with permission from Elsevier.

around the cutoff that would indicate manipulation of treatment assignment. Scatterplots of covariates against assignment can be used to check for continuity at the cutoff that would indicate whether units above and below the cutoff are indeed similar (57).

The RD estimator need not be interpreted only as the effect of a unit's exposure to the program (treatment) right at the cutoff value (47), but the assumption that units above and below the cutoff are similar except in their exposure to the program becomes less tenable as distance from the cutoff increases. Usual practice is to fit local linear regressions for observations within a narrow band on either side of the cutoff. Restricting the analysis in this way also means that nonlinearities in the relationship between the forcing and outcome variables are less important. One drawback is that smaller numbers of observations will yield less precise estimates, so the choice involves a trade-off between bias and precision.

The above approach works when the probability of a treatment jumps from 0 to 1 at the cutoff, which is known as sharp RD (**Figure 2*b***). If exposure is influenced by factors other than the value of the forcing variable, for example because administrators can exercise discretion over whom to include in the program or because individuals can, to some extent, manipulate their own assignment, the probability of treatment may take intermediate values close to the cutoff (**Figure 1*b***) and a modified approach known as fuzzy RD should be applied. This process uses the same two-stage approach to estimation as does an IV analysis (Models 5.1 and 5.2 in Appendix 1).

One example of a sharp RD design is Ludwig & Miller's (52) analysis of the US Head Start program. Help with applications for Head Start funding was targeted to counties with poverty rates of 59% or greater. This targeting led to a lasting imbalance in the receipt of Head Start funds among counties with poverty rates above and below the cutoff. Ludwig & Miller used local linear regressions of mortality on poverty rates for counties with poverty rates between 49% and 69%; the impact of the Head Start funding was defined as the difference between the estimated mortality rates at the upper and lower limits of this range. They found substantial reductions in mortality from causes amenable to Head Start but not from other causes of death or in children whose ages meant they were unlikely to benefit from the program.

Andalon (6) used a fuzzy RD design to investigate the impact of a conditional cash transfer program on obesity and overweight. Mexico's *Opportunidades* program provided substantial cash subsidies to households in rural communities that scored below a poverty threshold, which were conditional on school and health clinic attendance. There were a range of other ad hoc adjustments to eligibility criteria, creating a fuzzy rather than a sharp discontinuity in participation at the poverty cutoff. Andalon used two-stage least squares regression to estimate the effect of eligibility (based on the poverty score) on program participation and the effect of predicted participation on obesity and overweight. The author found no effect for men but a substantial reduction in obesity among women. Further testing indicated no bunching of poverty scores around the cutoff and no significant discontinuity at the cutoff in a range of covariates. Inclusion of the covariates in the outcome regressions had little effect on the estimates, further supporting the assumption of local randomization.

RD methods are widely regarded as the closest approximation of an observational study to an RCT (5), but their real value derives from their wide applicability to the evaluation of social programs for which eligibility is determined by a score on some form of continuous scale and also from their reliance on relatively weak, directly testable assumptions. One key shortcoming is that restricting the bandwidth to reduce bias results in a loss of precision (46, 73), and estimates that may hold over only a small segment of the whole population exposed to the intervention. This restriction to a subset of the population may not matter if the intervention is expected to affect outcomes locally, as in the case of a minimum legal drinking age or if the substantive focus of the study is on the effect of a small change in the assignment rule. It is more serious when the outcome of interest is the effect on the whole population.

STRENGTHENING INFERENCE IN NATURAL EXPERIMENTAL STUDIES

Causal inference can be strengthened in NE studies by the inclusion of additional design features alongside the principal method of effect estimation. Studies should be based on a clear theoretical understanding of how the intervention achieves its effects and the processes that determine exposure. Even if the observed effects are large and rapidly follow implementation, confidence in attributing them to the intervention can be markedly improved by a detailed consideration of alternative explanations.

Qualitative research can strengthen the design of RCTs of complex public health interventions (10, 56), and this argument applies equally to NEs (38). Qualitative research undertaken in preparation for, or alongside, NE studies can help to identify which outcomes might change as a consequence of the intervention and which are priorities for decision makers (42). It can also improve understanding of the processes that determine exposure, factors associated with intervention delivery and compliance, mechanisms by which outcomes are realized, and the strengths and limitations of routinely collected measures of exposures and outcomes (13). Qualitative studies conducted alongside the quantitative evaluation of Scotland's smoke-free legislation have been used to assess compliance with the intervention (24) and to identify a range of secondary outcomes such as changes in smoking behavior within the home (60). Qualitative methods for identifying the effects of interventions have also been proposed, but further studies are needed to establish their validity and usefulness (63, 72, 77).

Quantitative methods for strengthening inference include the use of multiple estimation methods within studies, replication studies, and falsification tests. Tests specific to particular methods, such as visual checks for discontinuities in RD and ITS studies, can also be used. Good-quality NE studies typically use a range of approaches. Comparing results obtained using different methods

can be used to assess the dependence of findings on particular assumptions (50). Such comparisons are particularly useful in early applications of novel methods whose strengths and weaknesses are not fully understood (49).

Falsification or placebo tests assess the plausibility of causal attribution by checking for the specificity of effects. One such approach is to use nonequivalent dependent variables to measure changes in outcomes that are not expected to respond to the intervention. They serve as indicators of residual confounding or the effects of other interventions introduced alongside the study intervention. A related approach is to use false implementation dates and to compare changes associated with those dates with effects estimated for the real implementation date. A similar test used in synthetic control studies involves generating placebo effects by replacing the intervention area with each of the areas in the donor pool in turn and then comparing the estimated intervention effect with the distribution of placebo effects (1, 2).

Most NE studies are conducted retrospectively, using data collected before the study is planned. Ideally, an analysis protocol, setting out hypotheses and methods, should be developed before any data analysis is conducted (21). Even when such protocols are published, they do not provide a perfect safeguard against selective reporting of positive findings. Replication studies, which by definition retest a previously published hypothesis, are a valuable additional safeguard against retrospectively fitting hypotheses to known features of the data. Reporting of NE studies of all kinds may also be improved by following established reporting guidelines such as STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) (74) or TREND (Transparent Reporting of Evaluations with Nonrandomized Designs) (28).

CONCLUSIONS

NE approaches to evaluation have become topical because they address researchers' and policy makers' interests in understanding the impact of large-scale population health interventions that, for practical, ethical, or political reasons, cannot be manipulated experimentally. We have suggested a pragmatic approach to NEs. NEs are not the answer to every evaluation question, and it is not always possible to conduct a good NE study whenever an RCT would be impractical. Choices among evaluation approaches are best made according to specific features of the intervention in question, such as the allocation process, the size of the population exposed, the availability of suitable comparators, and the nature of the expected impacts, rather than on the basis of general rules about which methods are strongest, regardless of circumstances. Availability of data also constrains the choice of methods. Where data allow, combining methods and comparing results are good ways to avoid overdependence on particular assumptions. Having a clear theory of change based on a sound qualitative understanding of the causal mechanisms at work is just as important as sophisticated analytical methods.

Many of the examples discussed above use routinely collected data on outcomes such as mortality, road traffic accidents, and hospital admissions and data on exposures such as poverty rates, alcohol sales, and tobacco consumption. Continued investment in such data sources, and in population health surveys, is essential if the potential for NEs to contribute to the evidence base for policy making is to be realized. Recent investments in infrastructure to link data across policy sectors for research purposes are a welcome move that should increase opportunities to evaluate NEs (12, 26, 37). Funding calls for population health research proposals should take a similarly even-handed approach to specifying which approaches would be acceptable and should emphasize the importance of developing a clear theory of change, carefully testing assumptions, and comparing estimates from alternative methods.

APPENDIX 1: STATISTICAL MODELS FOR NATURAL EXPERIMENTAL STUDIES

The standard multivariate regression model can be used to estimate the effect of exposure to an intervention (E) on the outcome (Y), with adjustment for a measured confounder (X),

$$Y_i = \beta_0 + \beta_1 E_i + \beta_2 X_i + \varepsilon_i. \qquad \text{(Model 1)}$$

Terms can be added to adjust for other confounders.

In a DiD analysis, observations are made on different units i at times t and the regression model includes an additional term for the period (P) in which the observation took place (coded 0 for preintervention or 1 for postintervention), and an interaction term between the period and exposure, which provides the effect estimate β_3:

$$Y_{it} = \beta_0 + \beta_1 E_i + \beta_2 P_t + \beta_3 E_i \times P_t + \varepsilon_{it}. \qquad \text{(Model 2)}$$

Segmented regression models estimate the baseline level of the outcome, the trend in the outcome before the intervention, the change that occurs at the point when the intervention is introduced, and the trend postintervention:

$$Y_t = \beta_0 + \beta_1 U_t + \beta_2 P_t + \beta_3 V_t + \varepsilon_t, \qquad \text{(Model 3)}$$

where U is the time from the start of the observation period, P is again a dummy variable indicating the preintervention ($P = 0$) and postintervention ($P = 1$) periods, and V is the time postintervention. Additional terms can be added to allow for multiple interventions, to model a lag period postintervention if it takes time for the effects to appear, or to account for serial correlation in the data.

In a sharp RD analysis, the model includes the forcing variable (Z), and the exposure variable (E) takes a value of 1 when Z is equal to or greater than the cutoff (C) that determines exposure and 0 otherwise. In Model 4, β_1 provides an estimate of the change in Y at the cutoff, and β_2 and β_3 estimate the slope below and above the cutoff. The analysis is usually restricted to values of Z close to C:

$$Y_i = \beta_0 + \beta_1 E_i + \beta_2 (1 - E_i)(Z_i - C) + \beta_3 E_i (Z_i - C) + \varepsilon_i. \qquad \text{(Model 4)}$$

Models 1–4 estimate an average treatment effect across the whole exposed population. The two-stage least squares (2SLS) models used in fuzzy RD and IV studies estimate a complier average causal effect, i.e., the effect of the intervention on those who comply with their assignment. The first stage predicts the probability of exposure π, with C representing the cutoff and Z the forcing variable in an RD analysis and the IV and a confounding variable in an IV analysis:

$$E_i \sim Bernoulli(\pi_i) \qquad \text{(Model 5.1)}$$

$$g(\pi_i) = \beta_0^* + \beta_1^* C_i + \beta_2^* Z_i,$$

where $g(.)$ denotes an appropriate link function, e.g., logit, probit.

The second stage uses the expected probabilities from the first stage to provide the effect estimate, β_1:

$$Y_i = \beta_0 + \beta_1 \hat{\pi}_i + \beta_2 Z_i + \varepsilon_i, \qquad \text{(Model 5.2)}$$

where $\hat{\pi}_i$ is the predicted probability of exposure from Model 5.1, e.g., for a logit link,

$$\hat{\pi}_i = \frac{1}{1 + \exp\left\{-\left(\hat{\beta}_0^* + \hat{\beta}_1^* C_i + \hat{\beta}_2^* Z_i\right)\right\}}.$$

Note that Model 5.2 includes the forcing variable and any other confounders from Model 5.1.

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

The authors receive core funding from the UK Medical Research Council (funding codes: MC_UU_12017/13, MC_UU_12017/15) and the Scottish Government Chief Scientist Office (funding codes: SPHSU13 and SPHSU15). In addition, S.V.K. is funded by an NHS Research Scotland Senior Clinical Fellowship (SCAF/15/02). The funders had no role in the preparation or submission of the manuscript, and the views expressed are those of the authors alone.

LITERATURE CITED

1. Abadie A, Diamond A, Hainmueller J. 2010. Synthetic control methods for comparative case studies: estimating the effect of California's Tobacco Control Program. *J. Am. Stat. Assoc.* 105:493–505
2. Abadie A, Diamond A, Hainmueller J. 2011. Synth: an R package for synthetic control methods in comparative case studies. *J. Stat. Softw.* 42:1–17
3. Abadie A, Diamond A, Hainmueller J. 2015. Comparative politics and the synthetic control method. *Am. J. Polit. Sci.* 50:495–510
4. Abadie A, Gardeazabal J. 2003. The economic costs of conflict: a case study of the Basque Country. *Am. Econ. Rev.* 93:113–32
5. Acad. Med. Sci. 2007. *Identifying the Environmental Causes of Disease: How Should We Decide What to Believe and When to Take Action.* London: Acad. Med. Sci.
6. Andalon M. 2011. Impact of Oportunidades in overweight and obesity in Mexico. *Health Econ.* 20(Suppl. 1):1–18
7. Austin PC. 2011. An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivar. Behav. Res.* 46:399–424
8. Basu S, Rehkopf DH, Siddiqi A, Glymour MM, Kawachi I. 2016. Health behaviors, mental health, and health care utilization among single mothers after welfare reforms in the 1990s. *Am. J. Epidemiol.* 83:531–38
9. Bauhoff S. 2014. The effect of school district nutrition policies on dietary intake and overweight: a synthetic control approach. *Econ. Hum. Biol.* 12:45–55
10. Bonell C, Fletcher A, Morton M, Lorenc T, Moore L. 2012. Realist randomised controlled trials: a new approach to evaluating complex public health interventions. *Soc. Sci. Med.* 75:2299–306
11. Bor J, Moscoe E, Mutevedzi P, Newell M-L, Barnighausen T. 2014. Regression discontinuity designs in epidemiology: causal inference without randomized trials. *Epidemiology* 25:729–37
12. Boyd J, Ferrante AM, O'Keefe C, Bass AJ, Randall AM, et al. 2012. Data linkage infrastructure for cross-jurisdictional health-related research in Australia. *BMC Health Serv. Res.* 2:480
13. Brown J, Neary J, Katikireddi SV, Thomson H, McQuaid RW, et al. 2015. Protocol for a mixed-methods longitudinal study to identify factors influencing return to work in the over 50s participating in the UK Work Programme: Supporting Older People into Employment (SOPIE). *BMJ Open* 5:e010525
14. Chattopadhyay R, Duflo E. 2004. Women as policy makers: evidence from a randomised policy experiment in India. *Econometrica* 72:1409–43
15. Comm. Soc. Determinants Health. 2008. *Closing the Gap in a Generation: Health Equity Through Action on the Social Determinants of Health. Final Report of the Commission on Social Determinants of Health.* Geneva: World Health Organ.
16. Craig P, Cooper C, Gunnell D, Macintyre S, Petticrew M, et al. 2012. Using natural experiments to evaluate population health interventions: new Medical Research Council guidance. *J. Epidemiol. Community Health* 66:1182–86

17. Crifasi CK, Meyers JS, Vernick JS, Webster DW. 2015. Effects of changes in permit-to-purchase handgun laws in Connecticut and Missouri on suicide rates. *Prev. Med.* 79:43–49
18. D'Agostino RB. 1998. Tutorial in biostatistics. Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. *Stat. Med.* 17:2265–81
19. De Angelo G, Hansen B. 2014. Life and death in the fast lane: police enforcement and traffic fatalities. *Am. Econ. J. Econ. Policy* 6:231–57
20. Deaton A. 2010. Instruments, randomisation and learning about development. *J. Econ. Lit.* 48:424–55
21. Dundas R, Ouédraogo S, Bond L, Briggs AH, Chalmers J, et al. 2014. Evaluation of health in pregnancy grants in Scotland: a protocol for a natural experiment. *BMJ Open* 4:e006547
22. Dunning T. 2012. *Natural Experiments in the Social Sciences: A Design-Based Approach*. Cambridge, UK: Cambridge Univ. Press
23. Dusheiko M, Gravelle H, Jacobs R, Smith P. 2006. The effect of financial incentives on gatekeeping doctors: evidence from a natural experiment. *J. Health Econ.* 25:449–78
24. Eadie D, Heim D, MacAskill S, Ross A, Hastings G, Davies J. 2008. A qualitative analysis of compliance with smoke-free legislation in community bars in Scotland: implications for public health. *Addiction* 103:1019–26
25. Fall T, Hägg S, Mägi R, Ploner A, Fischer K, et al. 2013. The role of adiposity in cardiometabolic traits: a Mendelian randomization analysis. *PLOS Med.* 10:e1001474
26. Farr Inst. Health Inf. Res. 2016. *Environmental and public health research*. Farr Inst. Health Inf. Res., Dundee, UK. **http://www.farrinstitute.org/research-education/research/environmental-and-public-health**
27. Foresight. 2007. *Tackling Obesities: Future Choices. Challenges for Research and Research Management.* London: Gov. Off. Sci.
28. Fuller T, Peters J, Pearson M, Anderson R. 2014. Impact of the transparent reporting of evaluations with nonrandomized designs reporting guideline: ten years on. *Am. J. Public Health* 104:e110–17
29. Goodman A, van Sluijs EMF, Ogilvie D. 2016. Impact of offering cycle training in schools upon cycling behaviour: a natural experimental study. *Int. J. Behav. Nutr. Phys. Act.* 13:34
30. Green CP, Heywood JS, Navarro M. 2014. Did liberalising bar hours decrease traffic accidents? *J. Health Econ.* 35:189–98
31. Grundy C, Steinbach R, Edwards P, Green J, Armstrong B, et al. 2009. Effect of 20 mph traffic speed zones on road injuries in London, 1986–2006: controlled interrupted time series analysis. *BMJ* 339:b4469
32. Gunnell D, Fernando R, Hewagama M, Priyangika W, Konradsen F, Eddleston M. 2007. The impact of pesticide regulations on suicide in Sri Lanka. *Int. J. Epidemiol.* 36:1235–42
33. Heckman JJ. 1995. Randomization as an instrumental variable. *Rev. Econ. Stat.* 78:336–41
34. Hernán M, Robins J. 2017. *Causal Inference*. Boca Raton, FL: Chapman Hall/CRC. In press
35. Hernán M, Robins JM. 2006. Instruments for causal inference. An epidemiologist's dream. *Epidemiology* 17:360–72
36. Holmes MV, Dale CE, Zuccolo L, Silverwood RJ, Guo Y, et al. 2014. Association between alcohol and cardiovascular disease: Mendelian randomisation analysis based on individual participant data. *BMJ* 349:g4164
37. House Commons Sci. Technol. Comm. 2016. *The Big Data Dilemma. Fourth Report of Session 2015–16.* HC 468. London: Station. Off. Ltd.
38. Humphreys DK, Panter J, Sahlqvist S, Goodman A, Ogilvie D. 2016. Changing the environment to improve population health: a framework for considering exposure in natural experimental studies. *J. Epidemiol. Community Health.* doi: 10.1136/jech-2015-206381
39. Ichida Y, Hirai H, Kondo K, Kawachi I, Takeda T, Endo H. 2013. Does social participation improve self-rated health in the older population? A quasi-experimental intervention study. *Soc. Sci. Med.* 94:83–90
40. IOM (Inst. Med.). 2010. *Bridging the Evidence Gap in Obesity Prevention: A Framework to Inform Decision Making*. Washington, DC: Natl. Acad. Press
41. Jones A, Rice N. 2009. *Econometric evaluation of health policies*. HEDG Work. Pap. 09/09. Univ. York
42. Katikireddi SV, Bond L, Hilton S. 2014. Changing policy framing as a deliberate strategy for public health advocacy: a qualitative policy case study of minimum unit pricing of alcohol. *Milbank Q.* 92:250–83

43. Katikireddi SV, Der G, Roberts C, Haw S. 2016. Has childhood smoking reduced following smoke-free public places legislation? A segmented regression analysis of cross-sectional UK school-based surveys. *Nicotine Tob. Res.* 18:1670–74
44. Kontopantelis E, Doran T, Springate DA, Buchan I, Reeves D. 2015. Regression based quasi-experimental approach when randomisation is not an option: interrupted time series analysis. *BMJ* 350:h2750
45. Kreif N, Grieve R, Hangartner D, Nikolova S, Turner AJ, Sutton M. 2015. Examination of the synthetic control method for evaluating health policies with multiple treated units. *Health Econ*. doi: 10.1002/hec.3258
46. Labrecque JA, Kaufman JS. 2016. Can a quasi-experimental design be a better idea than an experimental one? *Epidemiology* 27:500–2
47. Lee DS, Lemieux T. 2010. Regression discontinuity designs in economics. *J. Econ. Lit.* 48:281–355
48. Lewis SJ, Araya R, Davey Smith G, Freathy R, Gunnell D, et al. 2011. Smoking is associated with, but does not cause, depressed mood in pregnancy—a Mendelian randomization study. *PLOS ONE* 6:e21689
49. Linden A, Adams JL. 2011. Applying a propensity score-based weighting model to interrupted time series data: improving causal inference in programme evaluation. *J. Eval. Clin. Pract.* 17:1231–38
50. Linden A, Adams JL. 2012. Combining the regression discontinuity design and propensity score-based weighting to improve causal inference in program evaluation. *J. Eval. Clin. Pract.* 18:317–25
51. Little RJ, Rubin DB. 2000. Causal effects in clinical and epidemiological studies via potential outcomes: concepts and analytical approaches. *Annu. Rev. Public Health* 21:121–45
52. Ludwig J, Miller D. 2007. Does Head Start improve children's life chances? Evidence from an RD design. *Q. J. Econ.* 122:159–208
53. Mcleod AI, Vingilis ER. 2008. Power computations in time series analyses for traffic safety interventions. *Accid. Anal. Prev.* 40:1244–48
54. Melhuish E, Belsky J, Leyland AH, Barnes J, Natl. Eval. Sure Start Res. Team. 2008. Effects of fully-established Sure Start Local Programmes on 3-year-old children and their families living in England: a quasi-experimental observational study. *Lancet* 372:1641–47
55. Messer LC, Oakes JM, Mason S. 2010. Effects of socioeconomic and racial residential segregation on preterm birth: a cautionary tale of structural confounding. *Am. J. Epidemiol.* 171:664–73
56. Moore G, Audrey S, Barker M, Bond L, Bonell C, et al. 2015. MRC process evaluation of complex intervention. Medical Research Council guidance. *BMJ* 350:h1258
57. Moscoe E, Bor J, Barnighausen T. 2015. Regression discontinuity designs are under-used in medicine, epidemiology and public health: a review of current and best practice. *J. Clin. Epidemiol.* 68:132–43
58. Nandi A, Hajizadeh M, Harper S, Koski A, Strumpf EC, Heymann J. 2016. Increased duration of paid maternity leave lowers infant mortality in low and middle-income countries: a quasi-experimental study. *PLOS Med.* 13:e1001985
59. Pega F, Blakely T, Glymour MM, Carter KN, Kawachi I. 2016. Using marginal structural modelling to estimate the cumulative impact of an unconditional tax credit on self-rated health. *Am. J. Epidemiol.* 183:315–24
60. Phillips R, Amos A, Ritchie D, Cunningham-Burley S, Martin C. 2007. Smoking in the home after the smoke-free legislation in Scotland: qualitative study. *BMJ* 335:553
61. Ramsay CR, Matowe L, Grilli R, Grimshaw JM, Thomas RE. 2005. Interrupted time series designs in health technology assessment: lessons from two systematic reviews of behaviour change strategies. *Int. J. Technol. Assess. Health Care* 19:613–23
62. Restrepo BJ, Rieger M. 2016. Denmark's policy on artificial *trans* fat and cardiovascular disease. *Am. J. Prev. Med.* 50:69–76
63. Rihoux B, Ragin C. 2009. *Configurational Comparative Methods: Qualitative Comparative Analysis (QCA) and Related Techniques*. London: Sage
64. Robinson M, Geue C, Lewsey J, Mackay D, McCartney G, et al. 2014. Evaluating the impact of the Alcohol Act on off-trade alcohol sales: a natural experiment in Scotland. *Addiction* 109:2035–43
65. Rosenbaum PR, Rubin DB. 1983. The central role of the propensity score in observational studies for causal effects. *Biometrika* 70:41–55
66. Rubin DB. 2008. For objective causal inference, design trumps analysis. *Ann. Appl. Stat.* 2:808–40

67. Ryan AM, Krinsky S, Kontopantelis E, Doran T. 2016. Long-term evidence for the effect of pay-for-performance in primary care on mortality in the UK: a population study. *Lancet* 388:268–74
68. Sanson-Fisher RW, D'Este CS, Carey ML, Noble N, Paul CL. 2014. Evaluation of systems-oriented public health interventions: alternative research designs. *Annu. Rev. Public Health* 35:9–27
69. Shadish WR, Cook TD, Campbell DT. 2002. *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. New York: Houghton Mifflin
70. Shah BR, Laupacis A, Hux JE, Austin PC. 2005. Propensity score methods gave similar results to traditional regression modeling in observational studies: a systematic review. *J. Clin. Epidemiol.* 58:550–59
71. Swanson SA, Hernán MA. 2013. Commentary: how to report instrumental variable analyses (suggestions welcome). *Epidemiology* 24:370–74
72. Thomas J, O'Mara-Eves A, Brunton G. 2014. Using qualitative comparative analysis (QCA) in systematic reviews of complex interventions: a worked example. *Syst. Rev.* 3:67
73. van Leeuwen N, Lingsma HF, de Craen T, Nieboer D, Mooijaart S, et al. 2016. Regression discontinuity design: simulation and application in two cardiovascular trials with continuous outcomes. *Epidemiology* 27:503–11
74. von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, et al. 2008. Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. *J. Clin. Epidemiol.* 61:344–49
75. Wagner AK, Soumerai SB, Zhang F, Ross-Degnan D. 2002. Segmented regression analysis of interrupted time series studies in medication use research. *J. Clin. Pharm. Ther.* 27:299–309
76. Wanless D. 2004. *Securing Good Health for the Whole Population*. London: HM Treas.
77. Warren J, Wistow J, Bambra C. 2013. Applying qualitative comparative analysis (QCA) to evaluate a public health policy initiative in the North East of England. *Policy Soc.* 32:289–301
78. Yen ST, Andrews M, Chen Z, Eastwood DB. 2008. Food stamp program participation and food insecurity: an instrumental variables approach. *Am. J. Agric. Econ.* 90:117–32
79. Zhang F, Wagner AK, Ross-Degnan D. 2011. Simulation-based power calculation for designing interrupted time series analyses of health policy interventions. *J. Clin. Epidemiol.* 64:1252–61