Journal of Clinical Epidemiology 89 (2017) 30–42

Journal of Clinical Epidemiology

Quasi-experimental study designs series—paper 5: a checklist for classifying studies evaluating the effects on health interventions—a taxonomy without labels

Barnaby C. Reeves[a,*], George A. Wells[b], Hugh Waddington[c]

[a]*Clinical Trials and Evaluation Unit, School of Clinical Sciences, University of Bristol, Level 7 Queen's Building, Bristol Royal Infirmary, Bristol BS2 8HW, UK*
[b]*Department of Epidemiology and Community Medicine, Faculty of Medicine, University of Ottawa Heart Institute, 40 Ruskin Street, Ottawa, Ontario, Canada K1Y 4W7*
[c]*International Initiative for Impact Evaluation (3ie), 202-203, Rectangle One, D-4, Saket District Centre, New Delhi, 110017, India*

Accepted 6 February 2017; Published online 27 March 2017

Abstract

Objectives: The aim of the study was to extend a previously published checklist of study design features to include study designs often used by health systems researchers and economists. Our intention is to help review authors in any field to set eligibility criteria for studies to include in a systematic review that relate directly to the intrinsic strength of the studies in inferring causality. We also seek to clarify key equivalences and differences in terminology used by different research communities.

Study Design and Setting: Expert consensus meeting.

Results: The checklist comprises seven questions, each with a list of response items, addressing: clustering of an intervention as an aspect of allocation or due to the intrinsic nature of the delivery of the intervention; for whom, and when, outcome data are available; how the intervention effect was estimated; the principle underlying control for confounding; how groups were formed; the features of a study carried out after it was designed; and the variables measured before intervention.

Conclusion: The checklist clarifies the basis of credible quasi-experimental studies, reconciling different terminology used in different fields of investigation and facilitating communications across research communities. By applying the checklist, review authors' attention is also directed to the assumptions underpinning the methods for inferring causality.

Keywords: Health care; Health system; Evaluation; Study design; Quasi-experimental; Nonrandomized

1. Introduction

There are difficulties in drawing up a taxonomy of study designs to evaluate health care interventions or systems that do not use randomization [1]. To avoid the ambiguities of study design labels, a checklist of design features has been proposed by the Cochrane Non-Randomized Studies Methods Group (including B.C.R. and G.A.W.) to classify nonrandomized studies of health care interventions on the basis of what researchers did [1,2]. The checklist includes items about: whether a study made a comparison and, if yes, how comparison groups were formed; the timing of key elements of a study in relation to its conduct; and variables compared between intervention and comparator groups [1,2]. The checklist was created primarily from the perspective of health care evaluation, that is, the kinds of intervention most commonly considered in Cochrane reviews of interventions.

The checklist works well in principle for study designs in which the allocation mechanism applies to individual participants, although it does not characterize unit of analysis issues that may arise from the mechanism of allocation or the organizational hierarchy through which an intervention is provided (clustering by practitioner or organizational unit on which allocation is based), unit of treatment issues arising from the organizational hierarchy through which the intervention is provided, or unit of analysis issues arising from the unit at which data are collected and analysed (whether patient, practitioner or organisational aggregate). Most health interventions are delivered by discrete care provider units, typically organized hierarchically (e.g., hospitals, family practices, practitioners); this makes clustering important, except when allocation is randomized, because interventions are chosen by care provider units in complex

Funding: B.C.R is supported in part by the U.K. National Institute for Health Research Bristol Cardiovascular Biomedical Research Unit. H.W. is supported by 3ie.

* Corresponding author. Tel.: +4401173423143; fax: +4401173423288.
E-mail address: Barney.Reeves@bristol.ac.uk (B.C. Reeves).

http://dx.doi.org/10.1016/j.jclinepi.2017.02.016

What is new?

- Evaluations of health system interventions have features that differ and which are described differently compared to evaluations of health care interventions.
- An existing checklist of features has been extended to characterize: nesting of data in organizational clusters, for example, service providers; number of outcome measurements and whether outcomes were measured in the same or different individuals; whether the effects of an intervention are estimated by change over time or between groups; and the intrinsic ability of the analysis to control for confounding.
- Evaluations of health care and health system interventions have features that affect their credibility with respect to establishing causality but which are not captured by study design labels.
- With respect to inferring causality, review authors need to consider these features to discriminate ''strong'' from ''weak'' designs.
- Review authors can define eligibility criteria for a systematic review with reference to these study design features, but applying the checklist does not obviate the need for a careful risk of bias assessment.

ways. A modified checklist was also suggested for cluster-allocated designs (diverse study designs in which the allocation mechanism applies to groups of participants) [1,2], often used to evaluate interventions applied at the level of the group (e.g., disease prevention, health education, health policy), but the authors acknowledged that this checklist had not been well piloted.

There are three key challenges when trying to communicate study designs that do not use randomization to evaluate the effectiveness of interventions. First, study design labels are diverse or ambiguous, especially for cluster-allocated designs; moreover, there are key differences between research fields in the way that similar designs are conceived. Second, some study designs are, in fact, strategies for analysis rather than designs per se. Terms such as quasi-experimental, natural experiment, and observational cause particular ambiguity. The current checklist does not explicitly consider designs/analyses commonly used in health systems research (including so-called ''credible quasi-experimental studies'' [3,4]), often taking advantage of large administrative or other available data sets, and in other cases using data purposely collected as part of prospective designs where random assignment is not feasible. Third, and important with respect to the motivation for this paper, differences of opinion exist between health care and health systems researchers about the extent to which some studies are ''as good as'' randomized trials when well conducted; it is not clear whether this is because common designs are described with different labels or whether there are substantive differences. Therefore, our primary aim in this paper is revise the checklist to overcome these limitations.

Specific objectives were (1) to include a question to capture information about clustering; and (2) to extend the checklist to include study designs often used by health systems researchers and econometricians in a way that deals with the design/analysis challenge. We intended that the revised checklist should be able to resolve the differences in opinion about the extent to causality can be inferred from nonrandomized studies with different design features, improving communication between different health research communities. We did not intend that the checklist should be used as a tool to assess risk of bias, which can vary across studies with the same design features.

The paper is structured in three parts. Part 1 sets out designs currently used for health systems evaluations, illustrating their use through inclusion of different designs/analyses in a recent systematic review. Part 2 describes designs used for health intervention/program evaluations. Part 3 clarifies some of the ambiguities of study design labels using the proposed design feature framework.

2. Part 1: ''quasi-experimental'' studies considered by health system researchers and health economists

Health systems researchers and health economists use a wide range of ''quasi-experimental'' approaches to estimate causal effects of health care interventions. Some methods are considered stronger than others in estimating an unbiased causal relationship. ''Credible quasi-experimental studies'' are ones that ''estimate a causal relationship using exogenous variation in the exposure of interest which is not usually directly controlled the researcher.'' This exogenous variation refers to variation determined outside the system of relationships that are of interest and in some situations may be considered ''as good as random'' variation [3–5]. Credible quasi-experimental approaches are based on assignment to treatment and control that is not controlled by the investigators, and the term can be applied to different assignment rules; allocation to treatment and control is by definition not randomized, although some are based on identifying a source of variation in an exposure of interest that is assumed to be random (or exogenous). In the present context, they are considered to use rigorous designs and methods of analysis which can enable studies to adjust for unobservable sources of confounding [6] and are

identical to the union of "strong" and "weak" quasi-experiments as defined by Rockers et al. [4].

Credible quasi-experimental methods use assignment rules which are either known or can be modeled statistically, including: methods based on a threshold on a continuous scale (or ordinal scale with a minimum number of units) such as a test score (regression discontinuity design) or another form of "exogenous variation" arising, for example, due to geographical or administrative boundaries or assignment rules that have gone wrong (natural experiments). Quasi-experimental methods are also applied when assignment is self-selected by program administrators or by beneficiaries themselves [7,8]. Credible methods commonly used to identify causation among self-selected

Box 1 Thumbnail sketches of quasi-experimental studies used in program evaluations of CCT programs

Randomized controlled trial (RCT)	Individual participants, or clusters of participants, are randomly allocated to intervention or comparator.
Quasi-randomized controlled trial (Q-RCT)	Individual participants, or clusters of participants, are allocated to intervention or comparator in a quasi-random manner. For a credible study, the allocation mechanism should not be known to participants or any personnel responsible for data collection. The term natural experiment [4] is used instead when a study takes advantage of an "exogenous assignment" mechanism such as an error in implementation (as in the case of Morris et al. [9]), rather than explicit allocation by an experimenter or other decision maker who may be able to bias decisions about recruitment/participation.
Instrumental variable estimation (IVE)	Analysis of a cohort using an instrumental variable (IV) to estimate the effect of an intervention compared to a comparator in "two-stage" analysis. Requirements for a "good" IV are: (1) IV is strongly associated with allocation; (2) IV is independent of confounders between intervention and outcome; and (3) IV is independent of the outcome, given the allocation and confounders between allocation and the outcome [10].
Regression discontinuity (RD)	Analysis of a cohort which exploits local variation around a cutoff on a continuous "forcing" variable used by decision makers to determine allocation. A "good" forcing variable is: (1) strongly associated with allocation; (2) independent of confounders between intervention and outcome; and (3) independent of the outcome at the bandwidth around the cutoff.
Interrupted time series (ITS)	Analysis of a cohort with longitudinal "panel" data sets. In rare cases, the unit of analysis will be measured at the disaggregate level (i.e., the same people measured multiple times before and after treatment) [4]. Commonly, however, longitudinal data sets are clustered at aggregate levels of care (e.g., the health facility or district). In such cases, confounding by secular trends needs to be assessed, for example, with reference to a contemporaneous comparison group (controlled interrupted time series) and an assessment of performance bias—and some of the entries in the corresponding column in Table 2 would change.
Controlled interrupted time series (CITS)	As above for an interrupted time series but with data for a contemporaneous cohort with longitudinal "panel" data set for participants for whom the intervention is not implemented.
Difference study, including difference-in-differences study (DID)	Analysis of a cohort over time, in which no individuals have the intervention at the start and some receive the intervention by the end of the period of study. The typical study is clustered, with some clusters implementing the intervention; data are often also aggregated by cluster, for example, primary care practice. A "good" difference study is able to verify "common trends" and enables adjustment for probability of participation across groups (common support). A key feature of this design is the availability of longitudinal data for the same individuals for the entire period of study; studies that evaluate cluster-aggregated data often ignore changes in the individuals belonging to a cluster over time.
Cross-sectional study (XS)	The feature of this study design is that data required to classify individuals according to receipt of the intervention or comparator of interest and according to outcome are collected at the same time. Common methods of analysis include statistical matching (e.g., PSM) and adjusted regression analysis. A key limitation of this design is the inability to account for unobservable confounding and in some instances reverse causality.

groups include instrumental variable estimation (IVE), difference studies [including difference in differences, (DIDs)] and, to a lesser extent, propensity score matching (PSM) where individuals or groups are matched on preexisting characteristics measured at baseline and interrupted time series (ITS). Thumbnail sketches of these and other designs used by health system researchers are described in Box 1. It should be noted that the sketches of study types used by health program evaluators are not exhaustive. For example, pipeline studies, where treatment is withheld temporarily in one group until outcomes are measured (where time of treatment is not randomly allocated), are also used.

Quasi-experimental methods are used increasingly to evaluate programs in health systems research. Gaarder et al. [11], Baird et al. [12], and Kabeer and Waddington [13] have published reviews incorporating quasi-experimental studies on conditional cash transfer (CCT) programs, which make welfare benefits conditional upon beneficiaries taking specified actions like attending a health facility during the pre/post-natal period or enrolling children in school. Other reviews including quasi-experimental studies have evaluated health insurance schemes [14, 15] and maternal and child health programs [16]. Other papers in this themed issue of the Journal of Clinical Epidemiology describe how quasi-experimental studies can be identified for evidence synthesis [17], how data are best collected from quasi-experimental studies [18], and how the global capacity for including quasi-experimental studies in evidence synthesis can best be expanded [19, 20]. In this paper, we use studies from the reviews on the effects of CCT programs to illustrate the wide range of quasi-experimental methods used to quantify causal effects of the programs (Table 1).

Some of the earliest CCT programs randomly assigned clusters (communities of households) and used longitudinal household survey data collected by researchers to estimate the effects of CCTs on the health of both adults and children [21]. The design and analysis of a cluster-randomized controlled trial of this kind is familiar to health care researchers [29].

In other cases, it was not possible to assign beneficiaries randomly. In Jamaica's PATH program [22], benefits were allocated to people with scores below a criterion level on a multidimensional deprivation index and the effects of the program were estimated using a regression discontinuity analysis. This study involved recruiting a cohort of participants being considered for benefits, to whom a policy decision was applied (i.e., assign benefits or not on the basis the specified deprivation threshold). In such studies, by assigning the intervention on the basis of a cutoff value for a covariate, the assignment mechanism (usually correlated with the outcome of interest) is completely known and can provide a strong basis for inferences, although usually in a less efficient manner than in randomized controlled trials (RCTs). The treatment effect is estimated as the difference (''discontinuity'') between two predictions of the outcome based on the covariate (the average treatment effect at the cutoff): one for individuals just above the covariate cutoff (control group) and one for individuals just below the cutoff (intervention group) [30]. The covariate is often a test score (e.g., to decide who receives a health or education intervention) [31] but can also be distance from a geographic boundary [32]. Challenges of this design are assignment determined approximately, but not perfectly, by the cutoff [33] or circumstances in which participants may be able to control factors determining their assignment status such as their score or location.

As with health care evaluation, many studies in health systems research combine multiple methods. In Ecuador's Bono de Desarrollo Humano program, leakages in implementation caused ineligible families to receive the program, compromising the original discontinuity assignment. To compensate for this problem, the effects of the program were estimated as a ''fuzzy discontinuity'' using IVE [23]. An instrument (in this case, a dichotomous

Table 1. Experimental and quasi-experimental approaches applied in studies evaluating the effects of conditional cash transfer (CCT) programs

Study design label	Method of analysis	CCT program example
Randomized assignment	Bivariate (means comparison), multivariable regression	PROGRESSA, Mexico [21]
Regression discontinuity design	Regression analysis	Programme of Advancement Through Health and Education (PATH), Jamaica [22]
	Instrumental variables regression (''fuzzy'' discontinuity)	Bono de Desarrollo Humano (BDH), Ecuador [23]
Natural experiment	Instrumental variables (e.g., two-stage least squares) regression analysis	Bolsa Alimentação, Brazil [9]
Interrupted time series	Time-series regression analysis	Safe Delivery Incentive Programme (SDIP), Nepal [24]
Difference study	Difference-in-differences (DID) regression analysis	Familias en Accion, Colombia [25]
	Triple differences (DDD) regression analysis	Cambodia Education Sector Support Project (CESSP) [26]
Cohort study	Propensity score matching (PSM), retrospective cohort	Tekoporã, Paraguay [27]
Cross-sectional study	Propensity score matching (PSM), regression analysis	Bolsa Familia, Brazil [28]

Sources: reviews of CCTS by Gaarder et al. [11], Baird et al. [12] and Kabeer and Waddington [13].

variable taking the value of 1 or 0 depending on whether the participating family had a value on a proxy means test below or above a cutoff value used to determine eligibility to the program) must be associated with the assignment of interest, unrelated to potential confounding factors and related to the outcome of interest only by virtue of the relationship with the assignment of interest (and not, e.g., eligibility to another program which may affect the outcome of interest). If these conditions hold, then an unbiased effect of assignment can be estimated using two-stage regression methods [10]. The challenge lies not in the analysis itself (although such analyses are, typically, inefficient) but in demonstrating that the conditions for having a good instrument are met.

In the case of Bolsa Alimentação in Brazil, a computer error led eligible participants whose names contained nonstandard alphabetical characters to be excluded from the program. Because there are no reasons to believe that these individuals would have had systematically different characteristics to others, the exclusion of individuals was considered ''as good as random'' (i.e., a true natural experiment based on quasi-random assignment) [9].

Comparatively few studies in this review used ITS estimation, and we are not aware of any studies in this literature which have been able to draw on sufficiently long time series with longitudinal data for individual units of observation in order for the design to qualify ''as good as randomized.'' An evaluation of Nepal's Safe Delivery Incentive Programme (SDIP) drew on multiple cohorts of eligible households before and after implementation over a 7-year period [24]. The outcome (neonatal mortality) for each household was available at points in time that could be related to the inception of the program. Unfortunately, comparison group data were not available for nonparticipants, so an analysis of secular trends due to general improvements in maternal and child health care (i.e., not due to SDIP) was not possible. However, the authors were able to implement a regression ''placebo test'' (sometimes called a ''negative control''), in which SDIP treatment was linked to an outcome (use of antenatal care) which was not expected to be affected by the program, the rationale being that the lack of an estimated spike in antenatal care at the time of the expected change in mortality might suggest that these other confounding factors were not at play. But ultimately, due to the lack of comparison group data, the authors themselves note that the study is only able to provide ''plausible evidence of an impact'' rather than probabilistic evidence (p. 224).

Individual-level DID analyses use participant-level panel data (i.e., information collected in a consistent manner over time for a defined cohort of individuals). The Familias en Accion program in Colombia was evaluated using a DID analysis, where eligible and ineligible administrative clusters were matched initially using propensity scores. The effect of the intervention was estimated as the difference between groups of clusters that were or were not eligible for the intervention, taking into account the propensity scores on which they were matched [25]. DID analysis is only a credible method when we expect unobservable factors which determine outcomes to affect both groups equally over time (the ''common trends'' assumption). In the absence of common trends across groups, it is not possible to attribute the growth in the outcome to the program using the DID analysis. The problem is that we rarely have multiple period baseline data to compare variation between groups in outcomes over time before implementation, so the assumption is not usually verifiable. In such cases, placebo tests on outcomes which are related to possible confounders, but not the program of interest, can be investigated (see also above). Where multiple period baseline data are available, it may be possible to test for common trends directly and, where common trends in outcome levels are not supported, undertake a ''difference-in-difference-in-differences'' (DDDs) analysis. In Cambodia, the evaluators used DDD analysis to evaluate the Cambodia Education Sector Support Project, overcoming the observed lack of common trends in preprogram outcomes between beneficiaries and nonbeneficiaries [26].

As in the case of Attanasio et al. above [25], difference studies are usually made more credible when combined with methods of statistical matching because such studies are restricted to (or weighted by) individuals and groups with similar probabilities of participation based on observed characteristics—that is, observations ''in the region of common support.'' However, where panel or multiple time series cohort data are not available, statistical matching methods are often used alone. By contrast with the above examples, a conventional cohort study design was used to evaluate Tekoporã in Paraguay, relying on PSM and propensity weighted regression analysis of beneficiaries and nonbeneficiaries at entry into the cohort to control for confounding [27]. Similarly, for Bolsa Familia in Brazil evaluators applied PSM to cross-sectional (census) data [28]. Variables used to match observations in treatment and comparison should not be determined by program participation and are therefore best collected at baseline. However, this type of analysis alone does not satisfy the criterion of enabling adjustment for unobservable sources of confounding because it cannot rule out confounding of health outcomes data by unmeasured confounding factors, even when participants are well characterized at baseline.

3. Part 2: ''quasi-experimental'' designs used by health care evaluation researchers

The term ''quasi-experimental'' is also used by health care evaluation and social science researchers to describe studies in which assignment is nonrandom and influenced by the researchers. At the first appearance, many of the designs seem similar, although they are often labeled differently. Although an assignment rule may be known, it may

Box 2 Thumbnail sketches of quasi-experimental study designs used by health care evaluation researchers

Studies are cited which correspond to the way in which we conceive studies described with these labels.

Randomized controlled trial (RCT)	Individual participants, or clusters of participants, are randomly allocated to intervention or comparator. This design is the same as the RCT design described in Box 1.
Quasi-randomized controlled trial (Q-RCT)	Individual participants, or clusters of participants, are allocated to intervention or comparator in a quasi-random manner. In health care evaluation studies, the allocation rule is often by alternation, day of the week, odd/even hospital, or social security number [39]. The allocation rule may be as good as random but, typically, gives rise to a less credible study (compared to health system studies, where the allocation rule is applied by a higher level decision maker); if allocation is not concealed, research personnel who know the rule can recruit selectively or allocate participants in a biased way. This design is essentially the same as the Q-RCT design described in Box 1 but with different mechanisms for allocation.
Controlled before-and-after study (CBA)	Study in which outcomes are assessed at two time periods for several clusters (usually geographic). Clusters are classified into intervention and comparator groups. All clusters are studied without the intervention during period 1. Between periods 1 and 2, clusters in the intervention group implement the intervention of interest whereas clusters in the comparator group do not. The outcome for clusters receiving the intervention is compared to the outcome for comparator clusters during period 2, adjusted for the outcomes observed during period 1 (when no clusters had had the intervention). Observations usually represent episodes of care, so may or may not correspond to the same individuals during the two time periods. Data at either an aggregate [40] or individual level [41] can be analyzed. This design has similarities to the DID design described in Box 1.
Nonrandomized controlled trial (NRCT)	This is usually a prospective cohort study in which allocation to intervention and comparator is not random or quasi-random and is applied by research personnel [42]. The involvement of research personnel in the allocation rule may be difficult to discern; such studies may be labeled observational if the personnel responsible for the allocation rule are not clearly described or some personnel have both health care decision making and researcher roles. Individual-level data are usually analyzed. Note that nonrandom allocation of a health care intervention is often defined in relation to organizational factors (ward, clinic, doctor, provider organization) [43], and the analysis should take account of the data hierarchy if one exists.
Interrupted time series (ITS)	When used to study health care interventions, observations usually represent episodes of care or events, the cohorts studied may or may not correspond to the same individuals at different time points and are often clustered in organizational units (e.g., a health facility or district). (Such studies may be considered to consist of multiple cross-sectional ‘‘snapshots.’’) The analysis may be aggregated at the level of the clusters [44] or at the level of individual episodes of care [45]. If ITS do not have the benefit of analyzing multiple measurements from the same cohort over time (Box 1), confounding by secular trends needs to be assessed, for example, with reference to a contemporaneous comparison group (controlled interrupted time series, CITS, below). NB. Entries in Table 2 are for ITS as defined in Box 1; for ITS as defined here, entries for some cells would change. This design is similar to the ITS design described in Box 1.
Controlled interrupted time series (CITS)	As above for an ITS but with data for a contemporaneous comparison group in which the intervention was not implemented [46]. Measurements for the comparison group should be collected using the same methods. This design is similar to the CITS design described in Box 1.
Concurrently controlled prospective cohort study (PCS)	A cohort study in which subjects are identified prospectively and classified as having received the intervention or comparator of interest on the basis of the prospectively collected information [47]. Data for individuals are usually analyzed. However, it is important to note that nonrandom receipt of a health care intervention is almost always defined in relation to organizational factors (ward, clinic, doctor, provider organization), and the analysis should take into account the data hierarchy. This is equivalent to a ‘‘pipeline design’’ used in health systems program evaluation. It is very similar to a NRCT, except with respect to the method of allocation.
Concurrently controlled retrospective cohort study (RCS)	A cohort study in which subjects are identified from historic records and classified as having received the intervention or comparator of interest on the basis of the historic information [48]. As for a PCS, data for individuals are usually analyzed, but the analysis should take account of the data hierarchy.
Historically controlled cohort study (HCS)	This type of cohort study is a combination of an RCS (for one group, usually receiving the comparator) and a PCS (for the second group, usually receiving the intervention) [49]. Thus, the comparison between groups is not contemporaneous. The analysis should take into account the data hierarchy.

Case–control study (CC)	Consecutive individuals experiencing an outcome of interest are identified, preferably prospectively, from within a defined population (but for whom relevant data have not been collected) and form a group of "cases" [50]. Individuals, sometimes matched to the cases, who did not experience the outcome of interest are also identified from within the defined population and form the group of "controls." Data characterizing the intervention or comparator received in the past are collected retrospectively from existing records or by interviewing participants. The receipt of the intervention or comparator of interest is compared among cases and controls. If applicable, the analysis should take into account the data hierarchy.
Nested case–control study (NCC)	Individuals experiencing an outcome of interest are identified from within a defined cohort (for which some data have already been collected) and form a group of "cases." Individuals, often matched to the cases, who did not experience the outcome of interest are also identified from within the defined cohort and form the group of "controls" [51]. Additional data required for the study, characterizing the intervention or comparator received in the past, are collected retrospectively from existing records or by interviewing participants. The receipt of the intervention or comparator of interest is compared among cases and controls. If applicable, the analysis should take into account the data hierarchy.
Before after study (BA)	As for CBA but without data for a control group of clusters [52]. An uncontrolled comparison is made between frequencies of outcomes for the two time points. This term may also be applied to a study in which a cohort of individuals have the outcome (e.g., function, symptoms, or quality of life) measured before an intervention and after the intervention [53]. This type of study comprises a single "exposed" cohort [54] (often called a "case series"), with the outcome measured before and after exposure. If applicable, the analysis should take into account the data hierarchy.
Cross-sectional study (XS)	The feature of this study design is that information required to classify individuals according to receipt of the intervention or comparator of interest and according to outcome are collected at the same time, sometimes preventing researchers from knowing whether the intervention preceded the outcome [55]. In cross-sectional studies of health interventions, despite collecting data about the intervention/comparator and outcome at one point in time, the nature of the intervention and outcome may allow one to be confident about whether the intervention preceded the outcome. This design is similar to the XS design described in Box 1.

not be exploitable in the way described above for health system evaluations; for example, quasi-random allocation may be biased because of a lack of concealment, even when the allocation rule is "as good as random."

Researchers also use more conventional epidemiological designs, sometimes called observational, that exploit naturally occurring variation. Sometimes, the effects of interventions can be estimated in these cohorts using instrumental variables (prescribing preference; surgical volume; geographic variation, distance from health care facility), quantifying the effects of an intervention in a way that is considered to be unbiased [34–36]. Instrumental variable estimation using data from a randomized controlled trial to estimate the effect of treatment in the treated, when there is substantial nonadherence to the allocated intervention, is a particular instance of this approach [37,38].

Nonrandomized study design labels commonly used by health care evaluation researchers include: nonrandomized controlled trial, controlled before-and-after study (CBA), interrupted time series study (ITS; and CITS), prospective, retrospective or historically controlled cohort studies (PCS, RCS and HCS respectively), nested case–control study, case–control study, cross-sectional study, and before-after study. Thumbnail sketches of these study designs are given in Box 2. In addition, researchers sometimes report findings for uncontrolled cohorts or individuals ("case" series or reports), which only describe outcomes after an intervention [54]; these are not considered further because these studies do not collect data for an explicit comparator. It should be noted that these sketches are the authors' interpretations of the labels; studies that other researchers describe using these labels may not conform to these descriptions.

The designs can have diverse features, despite having the same label. Particular features are often chosen to address the logistical challenges of evaluating particular research questions and settings. Therefore, it is not possible to illustrate them with examples drawn from a single review as in part 1; instead, studies exemplifying each design are cited across a wide range of research questions and settings. The converse also occurs, that is, study design labels are often inconsistently applied. This can present great difficulties when trying to classify studies, for example, to describe eligibility for inclusion in a review. Relying on the study design labels used by primary researchers themselves to describe their studies can lead to serious misclassifications.

For some generic study designs, there are distinct study types. For example, a cohort study can study intervention

and comparator groups concurrently, with information about the intervention and comparator collected prospectively (PCS) or retrospectively (RCS), or study one group retrospectively and the other group prospectively (HCS). These different kinds of cohort study are conventionally distinguished according to the time when intervention and comparator groups are formed, in relation to the conception of the study. Some studies are sometimes incorrectly termed PCS, in our view, when data are collected prospectively, for example, for a clinical database, but when definitions of intervention and comparator required for the evaluation are applied retrospectively; in our view, this should be an RCS.

4. Part 3: study design features and their role in disambiguating study design labels

Some of the study designs described in parts 1 and 2 may seem similar, for example, DID and CBA, although they are labeled differently. Some other study design labels, for example, CITS/ITS, are used in both types of literature. In our view, these labels obscure some of the detailed features of the study designs that affect the robustness of causal attribution. Therefore, we have extended the checklist of features to highlight these differences. Where researchers use the same label to describe studies with subtly different features, we do not intend to imply that one or other use is incorrect; we merely wish to point out that studies referred to by the same labels may differ in ways that affect the robustness of an inference about the causal effect of the intervention of interest.

The checklist now includes seven questions (Table 2). The table also sets out our responses for the range of study designs as described in Boxes 1 and 2. The response ''possibly'' (P) is prevalent in the table, even given the descriptions in these boxes. We regard this as evidence of the ambiguity/inadequate specificity of the study design labels.

Question 1 is new and addresses the issue of clustering, either by design or through the organizational structure responsible for delivering the intervention (Box 3). This question avoids the need for separate checklists for designs based on assigning individual and clusters. A ''yes'' response can be given to more than one response item; the different types clustering may both occur in a single study and implicit clustering can occur an individually allocated nonrandomized study.

Question 1 in the checklist distinguishes individual allocation, cluster allocation (explicit clustering), and clustering due to the organizational hierarchy involved in the delivery of the interventions being compared (implicit clustering). Users should respond factually, that is, with respect to the presence of clustering, without making a judgment about the likely importance of clustering (degree of dependence between observations within clusters).

Questions 2–4 are also new, replacing the first question (''Was there a relevant comparison?'') in the original checklist [1,2]. These questions are designed to tease apart the nature of the research question and the basis for inferring causality.

Question 2 classifies studies according to the number of times outcome assessments were available. In each case, the response items distinguish whether or not the outcome is assessed in the same or different individuals at different times. Only one response item can be answered ''yes.''

Treatment effects can be estimated as changes over time or between groups. Question 3 aims to classify studies according to the parameter being estimated. Response items distinguish changes over time for the same or different individuals. Only one response item can be answered ''yes.''

Question 4 asks about the principle through which the primary researchers aimed to control for confounding. Three response items distinguish methods that:

a. control in principle for any confounding in the design, that is, by randomization, IVE, or regression discontinuity;
b. control in principle for time invariant unobserved confounding, that is, by comparing differences in outcome from baseline to end of study, using longitudinal/panel data for a constant cohort; or
c. control for confounding only by known and observed covariates (either by estimating treatment effects in ''adjusted'' statistical analyses or in the study design by restricting enrollment, matching and/or stratified sampling on known, and observed covariates).

The choice between these items (again, only one can be answered ''yes'') is key to understanding the basis for inferring causality.

Questions 5–7 are essentially the same as in the original checklist [1,2]. Question 5 asks about how groups (of individuals or clusters) were formed because treatment effects are most frequently estimated from between group comparisons. An additional response option, namely by a forcing variable, has been included to identify credible quasi-experimental studies that use an explicit rule for assignment based on a threshold for a variable measured on a continuous or ordinal scale or in relation to a spatial boundary. When answering ''yes'' to this item, the review author should also identify the nature of the variable by answering ''yes'' to another item. Possible assignment rules are identified: the action of researchers, time differences, location differences, health care decision makers/practitioners, policy makers, on the basis of the outcome, or some other process. Other, nonexperimental, study designs should be classified by the method of assignment (same list of variables) but without there being an explicit assignment rule.

Question 6 asks about important features of a study in relation to the timing of their implementation. Studies are

Table 2. Quasi-experimental taxonomy features checklist

	RCT	Q-RCT	IV	RD	CITS	ITS	DID	CBA	NRCT	PCS	RCS	HCT	NCC	CC	XS	BA
1. Was the intervention/comparator: (answer "yes" to more than 1 item, if applicable)																
Allocated to (provided for/ administered to/chosen by) individuals?	P	P	Y	Y	P	P	P	P	P	P	P	P	Y	Y	P	P
Allocated to (provided for/ administered to/chosen by) clusters of individuals?[a]	P	P	N	N	P	P	P	P	P	P	P	P	N	N	P	P
Clustered in the way it was provided (by practitioner or organizational unit)?[b]	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P
2. Were outcome data available: (answer "yes" to only 1 item)																
After intervention/comparator only (same individuals)?	P	P	P	P	N	N	N	N	P	P	P	P	Y	Y	Y	N
After intervention/comparator only (not all same individuals)?	N	N	N	N	P	P	N	P	P	P	P	P	N	N	N	P
Before (once) AND after intervention/comparator (same individuals)?	P	P	P	P	N	N	N	P	P	P	P	P	N	N	P	Y
Before (once) AND after intervention/comparator (not all same individuals)?	N	N	N	N	P	P	P	P	P	P	P	P	N	N	N	P
Multiple times before AND multiple times after intervention/ comparator (same individuals)?	P	P	P	P	P	P	P	P	P	P	P	P	N	N	P	P
Multiple times before AND multiple times after intervention/ comparator (not all same individuals)?	N	N	N	N	P	P	P	P	N	N	N	N	N	N	N	N
3. Was the intervention effect estimated by: (answer "yes" to only one item)																
Change over time (same individuals at different time points)?	N	N	N	N	N	Y	N	N	N	N	N	N	N	N	N	P
Change over time (not all same individuals at different time points)?	N	N	N	N	N	Y	N	N	N	N	N	N	N	N	N	P
Difference between groups[c] (of individuals or clusters receiving either intervention or comparator)?	Y	Y	Y	Y	Y	N	Y	Y	Y	Y	Y	Y	Y	Y	Y	N
4. Did the researchers aim to control for confounding (design or analysis) (answer "yes" to only one item)																
Using methods that control in principle for any confounding?	Y	Y	Y	Y	Y	Y	N	N	N	N	N	N	N	N	N	N
Using methods that control in principle for time-invariant unobserved confounding?	N	N	N	N	N	N	Y	Y	N	N	N	N	N	N	N	N
Using methods that control only for confounding by observed covariates?	P	P	P	P	P	P	P	P	Y	Y	Y	Y	Y	Y	Y	N
5. Were groups of individuals or clusters formed by (answer "yes" to more than one item, if applicable)[e]																
Randomization?	Y	N	N	N	N	na	N	N	N	N	N	N	N	N	N	na
Quasi-randomization?	N	Y	N	N	N	na	N	N	N	N	N	N	N	N	N	na

(Continued)

Table 2. Continued

	RCT	Q-RCT	IV	RD	CITS	ITS	DID	CBA	NRCT	PCS	RCS	HCT	NCC	CC	XS	BA
Explicit rule for allocation based on a threshold for a variable measured on a continuous or ordinal scale or boundary (in conjunction with identifying the variable dimension, below)?	N	N	Y	Y	N	na	N	N	N	N	N	N	N	N	N	na
Some other action of researchers?	N	N	P	P	P	na	N	N	Y	P	P	P	N	N	N	na
Time differences?	N	N	N	N	Y	na	N	N	N	N	N	Y	N	N	N	na
Location differences?	N	N	P	P	P	na	P	P	P	P	P	P	N	N	P	na
Health care decision makers/ practitioners?	N	N	P	P	P	na	P	P	P	P	P	P	N	N	P	na
Participants' preferences?	N	N	P	N	N	na	P	P	P	P	P	P	N	N	P	na
Policy maker	N	N	P	P	P	na	P	P	P	P	P	P	N	N	P	na
On the basis of outcome?[d]	N	N	N	N	N	na	N	N	N	N	N	N	Y	Y	N	na
Some other process? (specify)	N	N	P	P	P	na	P	P	P	P	P	P	N	N	P	na
6. Were the following features of the study carried out after the study was designed (answer ''yes'' to more than one item, if applicable)																
Characterization of individuals/ clusters before intervention?	Y	Y	P	P	P	P	P	P	Y	Y	P	P	N	N	N	P
Actions/choices leading to an individual/cluster becoming a member of a group?[d]	Y	Y	P	P	P	na	P	P	Y	Y	P	P	N	N	N	na
Assessment of outcomes?	Y	Y	P	P	P	P	P	P	Y	Y	P	P	P	P	N	P
7. Were the following variables measured before intervention: (answer ''yes'' to more than one item, if applicable)																
Potential confounders?	P	P	P	P	P	N	P	P	P	P	P	P	P	P	N	N
Outcome variable(s)?	P	P	P	P	Y	Y	Y	Y	P	P	P	P	N	N	N	P

Abbreviations: RCT, randomized controlled trial; Q-RCT, quasi-randomized controlled trial; IV, instrumental variable; RD, regression discontinuity; CITS, controlled interrupted time series; ITS, interrupted time series; DID, difference-in-difference; CBA, controlled before-and-after study; NRCT, nonrandomized controlled trial; PCS, prospective cohort study; RCS, retrospective cohort study; HCT, historically controlled study; NCC, nested case–control study; CC, case–control study; XS, cross-sectional study; BA, before-after study; Y, yes; N, no; P, possibly; na, not applicable.

Cells in the table are completed with respect to the thumbnail sketches of the corresponding designs described in Boxes 1 and 2.

[a] This row describes ''explicit'' clustering. In randomized controlled trials, participants can be allocated individually or by virtue of ''belonging'' to a cluster such as a primary care practice or a village.

[b] This row describes ''implicit'' clustering. In randomized controlled trials, participants can be allocated individually but with the intervention being delivered in clusters (e.g., group cognitive therapy); similarly, in a cluster-randomized trial (by general practice), the provision of an intervention could also be clustered by therapist, with several therapists providing ''group'' therapy.

[c] A study should be classified as ''yes'' for this feature, even if it involves comparing the extent of change over time between groups.

[d] For (nested) case–control studies, group refers to the case/control status of an individual.

[e] The distinction between these options is to do with the exogeneity of the allocation, hence designs further to the right in the table are more to have involve allocation by some non-exogenous agent.

classified according to whether three key steps were carried out after the study was designed, namely: acquisition of source data to characterize individuals/clusters before intervention; actions or choices leading to an individual or cluster becoming a member of a group; and the assessment of outcomes. One or more of these items can be answered ''yes,'' as would be the case for all steps in a conventional RCT.

Question 7 asks about the variables that were measured and available to control for confounding in the analysis. The two broad classes of variables that are important are the identification and collection of potential confounder variables and baseline assessment of the outcome variable(s). The answers to this question will be less important if the researchers of the original study used a method to control for any confounding, that is, used a credible quasi-experimental design.

The health care evaluation community has historically been much more difficult to win around to the potential value of nonrandomized studies to evaluate interventions. We think that the checklist helps to explain why, that is, because designs used in health care evaluation do not often control for unobservables when the study features are examined carefully. To the extent that these features are immutable, the skepticism is justified. However, to the extent that studies may be possible with features that promote the credibility of causal inference, health care evaluation researchers may be missing an opportunity to provide high-quality evidence.

Reflecting on the circumstances of nonrandomized evaluations of health care and health system interventions may provide some insights why these different groups have disagreed

Box 3 Clustering in studies evaluating the effects of health system or health care interventions

Clustering is a potentially important consideration in both RCTs and nonrandomized studies. Clusters exist when observations are nested within higher level organizational units or structures for implementing an intervention or data collected; typically, observations within clusters will be more similar with respect to outcomes of interest than observations between clusters. Clustering is a natural consequence of many methods of nonrandomized assignment/designation because of the way in which many interventions are implemented. Analyses of clustered data that do not take clustering into account will tend to overestimate the precision of effect estimates.

Clustering occurs when implementation of an intervention is explicitly at the level of a cluster/organizational unit (as in a cluster-randomized controlled trial, in which each cluster is explicitly allocated to control or intervention). Clustering can also arise implicitly, from naturally occurring hierarchies in the data set being analyzed, that reflect clusters that are intrinsically involved in the delivery of the intervention or comparator. Both explicit and implicit clustering can be present in a single study.

Examples of types of cluster

- Practitioner (surgeon; therapist, family doctor; teacher; social worker; probation officer; etc.).
- Organizational unit [general practice, hospital (ward), community care team; school, etc.].
- Social unit (family unit; network of individuals clustered in some nongeographic network, etc.).
- Geographic area (health region; city jurisdiction; small electoral district, etc.).

"Explicit" clustering

- Clustering arising from allocation/formation of groups; clusters can contain only intervention or control observations.

"Implicit" clustering

- Clustering arising from naturally occurring hierarchies of units of analysis in the data set being analyzed to answer the research question.
- Clusters can contain intervention and control observations in varying proportions.
- Factors associated with designation as intervention or control may vary by cluster.

No clustering

- Designation of an observation as intervention or control is only influenced by the characteristics of the observation (e.g., patient choice to self-medicate with an over-the-counter medication; natural experiment in which allocation of individuals is effectively random, as in the case of Bolsa Alimentação where a computer error led to the allocation to intervention or comparator [31].)

about the credibility of effects estimated in quasi-experimental studies. The checklist shows that credible quasi-experimental studies gain credibility from using high-quality longitudinal/panel data; such data characterizing health care are rare, leading to evaluations that "make do" with the data that are available in existing information systems.

The risk of confounding in health care settings is inherently greater because participants' characteristics are fundamental to choices about interventions in usual care; mitigating against this risk requires high-quality clinical data to characterize participants at baseline and, for pharmaco-epidemiological studies about safety, often over time. Important questions about health care for which quasi-experimental methods of evaluation are typically considered are often to do with the outcome of discrete episodes of care, usually binary, rather than long-term outcomes for a cohort of individuals; this can lead to a focus on the invariant nature of the organizations providing the care rather than the varying nature of the individuals receiving care. These contrasts are apparent between, for example: DID studies using panel data to evaluate an intervention such as CCT among individuals with CBA studies of an intervention implemented at an organizational level studying multiple cross-sections of health care episodes; or credible and less credible interrupted time series.

There is a new article in the field of hospital epidemiology which also highlights various features of what it terms as quasi-experimental designs [56]. The list of features appears to be aimed at researchers designing a quasi-experimental study, acting more as a prompt (e.g., "consider options for …") rather than as a checklist for a researcher appraising a study to communicate clearly to others about the nature of a published study, which is our perspective (e.g., a review author). There is some overlap with our checklist, but the list described also includes several study attributes intended to reduce the risk of bias, for example,

blinding. By contrast, we consider that an assessment of the risk of bias in a study is essential and needs to be carried out as a separate task.

5. Conclusion

The primary intention of the checklist is to help review authors to set eligibility criteria for studies to include in a review that relate directly to the intrinsic strength of the studies in inferring causality. The checklist should also illuminate the debate between researchers in different fields about the strength of studies with different features—a debate which has to date been somewhat obscured by the use of different terminology by researchers working in different fields of investigation. Furthermore, where disagreements persist, the checklist should allow researchers to inspect the basis for these differences, for example, the principle through which researchers aimed to control for confounding and shift their attention to clarifying the basis for their respective responses for particular items.

Acknowledgments

Authors' contributions: All three authors collaborated to draw up the extended checklist. G.A.W. prepared the first draft of the paper. H.W. contributed text for Part 1. B.C.R. revised the first draft and created the current structure. All three authors approved submission of the final manuscript.

References

[1] Reeves BC, Deeks JJ, Higgins JPT, Wells GA. Chapter 13: including non-randomized studies. In: Higgins JPT, Green S (editors), Cochrane handbook for systematic reviews of interventions. Version 5.0.2 (updated September 2009). The Cochrane Collaboration, 2009. Available at www.cochrane-handbook.org.

[2] Higgins JPT, Ramsay C, Reeves BC, Shea B, Valentine J, Tugwell P, et al. Issues relating to study design and risk of bias when including non-randomized studies in systematic reviews on the effects of interventions. Res Synth Methods 2013;4:12–25.

[3] Rockers PC, Røttingen JA, Shemilt I, Tugwell P, Bärnighausen T. Inclusion of quasi-experimental studies in systematic reviews of health systems research. Health Policy 2015;19:511–21.

[4] Bärnighausen T, Tugwell P, Røttingen JA, Shemilt I, Rockers P, Geldsetzer P, et al. Quasi-experimental study designs series - Paper 4: uses and value. J Clin Epidemiol 2017;89:21–9.

[5] Bärnighausen T, Oldenburg C, Tugwell P, Bommer C, Cara Ebert, Barreto M, et al. Quasi-experimental study designs series - Paper 7: assessing the assumptions. J Clin Epidemiol 2017;89:53–66.

[6] Waddington H, Aloe AM, Becker BJ, Djimeu EW, Hombrados JG, Tugwell P, et al. Quasi-experimental study designs series-paper 6: risk of bias assessment. J Clin Epidemiol 2017;43–52.

[7] Cook TD, Campbell DT. Quasi-experimentation: design and analysis issues for field settings. Boston: Houghton Mifflin; 1979.

[8] Shadish WR, Cook TD, Campbell DT. Experimental and quasi-experimental designs for generalized causal inference. Boston, MA: Houghton Mifflin; 2002.

[9] Morris SS, Olinto P, Flores R, Nilson EA, Figueiro AC. Conditional cash transfers are associated with a small reduction in the rate of weight gain of preschool children in northeast Brazil. J Nutr 2004;134:2336–41.

[10] Greenland S. An introduction to instrumental variables for epidemiologists. Int J Epidemiol 2000;29:722–9.

[11] Gaarder MM, Glassman A, Todd JE. Conditional cash transfer programmes: opening the black box. Conditional cash transfers and health: unpacking the causal chain. J Development Effectiveness 2010;2:6–50.

[12] Baird S, Ferreira FHG, Özler B, Woolcock M. Relative effectiveness of conditional and unconditional cash transfers for schooling outcomes in developing countries: a systematic review. Campbell Syst Rev 2013;8. http://dx.doi.org/10.4073/csr.2013.8.

[13] Kabeer N, Waddington H. Economic impacts of conditional cash transfer programmes: a systematic review and meta-analysis. J Development Effectiveness 2015;7:290–303.

[14] Acharya A, Vellakkal S, Taylor F, Masset E, Satija A, Burke M, et al. Impact of national health insurance for the poor and the informal sector in low and middle-income countries: a systematic review. London: EPPI-Centre, Social Science Research Unit, Institute of Education, University of London. 2012; 2012.

[15] Giedion U, Alfonso EA, Díaz Y. The impact of universal coverage schemes in the developing world: a review of the existing evidence. UNICO Studies Series 25. Washington DC: The World Bank; 2013.

[16] Tanner JC, Aguilar Rivera AM, Candland T, Galdo V, Manang F, Trichler R, et al. Delivering the millennium development goals to reduce maternal and child mortality. A systematic review of impact evaluation evidence. World Bank; Washington DC: Independent Evaluation Group; 2013. Available at: https://ieg.worldbankgroup.org/Data/Evaluation/files/mch_eval_updated2.pdf. Accessed April 12, 2017.

[17] Glanville J, Eyers J, Jones AM, Shemilt I, Wang G, Johansen M, et al. Quasi-experimental study designs series-paper 8: identifying quasi-experimental studies to inform systematic reviews. J Clin Epidemiol 2017;89:67–76.

[18] Aloe AM, Becker BJ, Duvendack M, Valentine JC, Shemilt I, Waddington H. Quasi-experimental study designs series-paper 9: collecting data from quasi-experimental studies. J Clin Epidemiol 2017;89:77–83.

[19] Lavis JN, Bärnighausen T, El-Jardali F. Quasi-experimental study designs series - Paper 11: Supporting the production and use of health systems research syntheses that draw on quasi-experimental study designs. J Clin Epidemiol 2017;89:92–7.

[20] Rockers PC, Tugwell P, Grimshaw J, Oliver S, Atun R, Røttingen JA, et al. Quasi-experimental study designs series-paper 12: strengthening global capacity for evidence synthesis of quasi-experimental health systems research. J Clin Epidemiol 2017;89:98–105.

[21] Gertler PJ, Boyce S. An Experiment in Incentive-Based Welfare: The Impact of PROGESA on Health in Mexico 2001. Available at https://web.warwick.ac.uk/res2003/papers/Gertler.pdf.

[22] Levy D, Ohls J. Evaluation of Jamaica's PATH Program: Final Report. Washington, DC: Mathematica Policy Research, Inc; 2007.

[23] Oosterbeek H, Ponce J, Schady N. The Impact of cash transfers on school enrolment: evidence from Ecuador. World Bank; 2008: Policy Research Working Paper No. 4645. Available at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1149578.

[24] Powell-Jackson T, Neupane BD, Tiwari S, Tumbahangphe K, Manandhar D, Costello AM. The impact of Nepal's national incentive programme to promote safe delivery in the district of Makwanpur. Adv Health Econ Health Serv Res 2009;21:221–49.

[25] Attanasio O, Battistin E, Fitzsimons E, Mesnard A, Vera-Hernández M. How effective are conditional cash transfers? Evidence from Columbia. The Institute of Fiscal Studies; 2005: Briefing Note number 54. Available at http://www.ifs.org.uk/bns/bn54.pdf.

[26] Filmer D, Schady N. Getting girls into school: evidence from a scholarship program in Cambodia. Econ Development Cult Change 2008;56:581−617.
[27] Soares FV, Ribas RP, Hirata GI. Achievements and shortfalls of conditional cash transfers: impact evaluation of Paraguay's Tekoporã Programme. Publications 3, International Policy Centre for Inclusive Growth; 2008. Available at https://www.unicef.org/socialpolicy/files/Achievement_and_shortfalls_of_conditional_cash_transfers.pdf.
[28] Cardoso E, Souza AP. The impact of cash transfers on child labor and school enrollment in Brazil. Vanderbilt University; 2003: Working Paper # 04−W07.
[29] Ukuomonne OC, Gulliford MC, Chinn S, Sterne JAC, Burney PGJ. Methods for evaluating area-wide and organisation-based interventions in health and health care: a systematic review. Health Technol Assess 1999;3:1−92.
[30] Lee DS, Lemieux T. Regression discontinuity designs in economics. J Econ Lit 2010;48:281−355.
[31] Moss B, Yeaton W. Shaping policies related to developmental education: an evaluation using the regression-discontinuity design. Educ Eval Policy Anal 2006;28:215−29.
[32] Arcand JL, Djimeu Wouabe E. Teacher training and HIV/AIDS prevention in West Africa: regression discontinuity design evidence from the Cameroon. Health Econ 2010;19:36−54.
[33] Valentine JC, Thompson SG. Issues relating to the inclusion of nonrandomized studies in systematic reviews on the effects of interventions. Res Synth Methods 2013;4:26−35.
[34] Boef AG, Souverein PC, Vandenbroucke JP, van Hylckama Vlieg A, de Boer A, le Cessie S, et al. Instrumental variable analysis as a complementary analysis in studies of adverse effects: venous thromboembolism and second-generation versus third-generation oral contraceptives. Pharmacoepidemiol Drug Saf 2016;25(3):317−24.
[35] Pezzin LE, Laud P, Yen TW, Neuner J, Nattinger AB. Reexamining the relationship of breast cancer hospital and surgical volume to mortality: an instrumental variable analysis. Med Care 2015;53:1033−9.
[36] Bekelis K, Missios S, Coy S, Singer RJ, MacKenzie TA. New York state: comparison of treatment outcomes for unruptured cerebral aneurysms using an instrumental variable analysis. J Am Heart Assoc 2015;4(7). http://dx.doi.org/10.1161/JAHA.115.002190.
[37] Goldsmith LP, Lewis SW, Dunn G, Bentall RP. Psychological treatments for early psychosis can be beneficial or harmful, depending on the therapeutic alliance: an instrumental variable analysis. Psychol Med 2015;45:2365−73.
[38] Reeves BC, Pike K, Rogers CA, Brierley RC, Stokes EA, Wordsworth S, et al. A multicentre randomised controlled trial of Transfusion Indication Threshold Reduction on transfusion rates, morbidity and health-care resource use following cardiac surgery (TITRe2). Health Technol Assess 2016;20:1−260.
[39] Aiken AM, Davey C, Hargreaves JR, Hayes RJ. Re-analysis of health and educational impacts of a school-based deworming programme in western Kenya: a pure replication. Int J Epidemiol 2015;44:1572−80.
[40] Holloway KA, Gautam BR, Reeves BC. The effects of different kinds of user fees on prescribing quality in rural Nepal. J Clin Epidemiol 2001;54:1065−71.
[41] Collin S, Reeves BC, Hendy J, Fulop N, Hutchings A, Priedane E. Computerised physician order entry (CPOE) and picture archiving and communication systems (PACS) implementation in the NHS: a quantitative before-and-after study. BMJ 2008;337:a939.
[42] Carey ME, Mandalia PK, Daly H, Gray LJ, Hale R, Martin Stacey L, et al. Increasing capacity to deliver diabetes self-management education: results of the DESMOND lay educator non-randomized controlled equivalence trial. Diabet Med 2014;31:1431−8.
[43] Skre I, Friborg O, Breivik C, Johnsen LI, Arnesen Y, Wang CE. A school intervention for mental health literacy in adolescents: effects of a non-randomized cluster controlled trial. BMC Public Health 2013;13:873.
[44] Campbell SM, Reeves D, Kontopantelis E, Sibbald B, Roland M. Effects of pay for performance on the quality of primary care in England. N Engl J Med 2009;361:368−78.
[45] Grijalva CG, Nuorti JP, Arbogast PG, Martin SW, Edwards KM, Griffin MR. Decline in pneumonia admissions after routine childhood immunisation with pneumococcal conjugate vaccine in the USA: a time-series analysis. Lancet 2007;369:1179−86.
[46] Steinbach R, Perkins C, Tompson L, Johnson S, Armstrong B, Green J, et al. The effect of reduced street lighting on road casualties and crime in England and Wales: controlled interrupted time series analysis. J Epidemiol Community Health 2015;69:1118−24.
[47] Langham J, Reeves BC, Lindsay KW, van der Meulen JH, Kirkpatrick PJ, Gholkar AR, et al. For the steering group for national study of subarachnoid haemorrhage. Variation in outcome after subarachnoid haemorrhage. A study of neurosurgical units in UK and Ireland. Stroke 2009;40:111−8.
[48] Murphy GJ, Reeves BC, Rogers CA, Rizvi SIA, Culliford L, Angelini GD. Increased mortality, post-operative morbidity and cost following red blood cell transfusion in patients having cardiac surgery. Circulation 2007;116:2544−52.
[49] Sacks HS, Chalmers TC, Smith H. Randomized versus historical controls for clinical trials. Am J Med 1982;72:233−9.
[50] Vučković BA, van Rein N, Cannegieter SC, Rosendaal FR, Lijfering WM. Vitamin supplementation on the risk of venous thrombosis: results from the MEGA case-control study. Am J Clin Nutr 2015;101:606−12.
[51] Graham DJ, Campen D, Hui R, Spence M, Cheetham C, Levy G, et al. Risk of acute myocardial infarction and sudden cardiac death in patients treated with cyclo-oxygenase 2 selective and non-selective non-steroidal anti-inflammatory drugs: nested case-control study. Lancet 2005;365:475−81.
[52] Maini R, Van den Bergh R, van Griensven J, Tayler-Smith K, Ousley J, Carter D, et al. Picking up the bill - improving health-care utilisation in the Democratic Republic of Congo through user fee subsidisation: a before and after study. BMC Health Serv Res 2014;14:504.
[53] Hopkins C, Browne JP, Slack R, Lund V, Topham J, Reeves B, et al. The national comparative audit of surgery for nasal polyposis and chronic rhinosinusitis. Clin Otolaryngol 2006;31:390−8.
[54] Dekkers OM, Egger M, Altman DG, Vandenbroucke JP. Distinguishing case series from cohort studies. Ann Intern Med 2012;156:37−40.
[55] Agot KE, Ndinya-Achola JO, Kreiss JK, Weiss NS. Risk of HIV-1 in rural Kenya: a comparison of circumcised and uncircumcised men. Epidemiology 2004;15:157−63.
[56] Schweizer ML, Braun BI, Milstone AM. Research methods in healthcare epidemiology and antimicrobial stewardship—quasi-experimental designs. Infect Control Hosp Epidemiol 2016;37:1135−40. Published online: 07 June 2016.