ELSEVIER

CrossMark

Journal of Clinical Epidemiology 89 (2017) 43–52

Journal of Clinical Epidemiology

Quasi-experimental study designs series—paper 6: risk of bias assessment

Hugh Waddington[a,*], Ariel M. Aloe[b], Betsy Jane Becker[c], Eric W. Djimeu[a], Jorge Garcia Hombrados[d], Peter Tugwell[e], George Wells[e], Barney Reeves[f]

[a]*International Initiative for Impact Evaluation, New Delhi, India*
[b]*University of Iowa, Iowa City, IA, USA*
[c]*Florida State University, Tallahassee, FL, USA*
[d]*University of Sussex, Brighton, UK*
[e]*Department of Medicine, University of Ottawa, Ottawa, Canada*
[f]*University of Bristoln, Bristol, UK*

Accepted 6 February 2017; Published online 27 March 2017

Abstract

Objectives: Rigorous and transparent bias assessment is a core component of high-quality systematic reviews. We assess modifications to existing risk of bias approaches to incorporate rigorous quasi-experimental approaches with selection on unobservables. These are nonrandomized studies using design-based approaches to control for unobservable sources of confounding such as difference studies, instrumental variables, interrupted time series, natural experiments, and regression-discontinuity designs.

Study Design and Setting: We review existing risk of bias tools. Drawing on these tools, we present domains of bias and suggest directions for evaluation questions.

Results: The review suggests that existing risk of bias tools provide, to different degrees, incomplete transparent criteria to assess the validity of these designs. The paper then presents an approach to evaluating the internal validity of quasi-experiments with selection on unobservables.

Conclusion: We conclude that tools for nonrandomized studies of interventions need to be further developed to incorporate evaluation questions for quasi-experiments with selection on unobservables.

Keywords: Risk of bias; Systematic review; Meta-Analysis; Quasi-experiment; Natural experiment; Instrumental variables; Regression discontinuity; Interrupted time series; Difference in differences

1. Introduction

Researchers in health and the social sciences quantify treatment effects—that is, changes in outcomes which are attributed to a particular intervention—using a range of nonrandomized approaches, also called quasi-experiments (QEs) [1–3]. QEs are quantitative studies which are used to make causal inferences when treatment is by definition not randomly assigned. There are two main types of QE study: designs which are able to adjust for unobservable sources of confounding (“selection on unobservables”); and methods which adjust for observables directly (e.g., analysis of variance or adjusted regression analysis) whose validity is based on the assumption of unconfoundedness [4,5]. In this paper, we discuss explicitly approaches to control for selection on unobservables, including difference in differences (DID), instrumental variables, interrupted time series (ITS), natural experiments, and regression discontinuity designs. Often these designs are combined with methods to control for observable confounding such as statistical matching [e.g., propensity score matching (PSM)].

All quantitative causal studies are subject to biases relating to design (internal validity) and methods of statistical analysis (statistical conclusion validity) [3]. In the same way that experimental studies [randomized controlled trials (RCTs) can have methodological problems in implementation (e.g., contamination of controls, poor allocation concealment, nonrandom attrition, and so on), inappropriately designed or executed QEs will not generate good causal evidence. QE studies are, however, potentially at higher risk of bias than their experimental counterparts [6,7], with perhaps the most critical biases for causal

* Corresponding author. Tel.: +44-7779-261108; fax: +44-2030-738303.
E-mail address: hwaddington@3ieimpact.org (H. Waddington).

http://dx.doi.org/10.1016/j.jclinepi.2017.02.015
0895-4356/

What is new?

Key findings

- Rigorous nonrandomized studies use design-based approaches which can control for unobservable sources of confounding. These include difference studies, instrumental variables estimation, interrupted time series, natural experiments, and regression discontinuity designs. Systematic critical appraisal of these studies requires identification of the design and assessment of the methodology, which existing risk of bias tools can incorporate.

What this adds to what was known?

- A review of risk of bias tools suggests that they provide, to different degrees, incomplete transparent criteria to assess rigorous nonrandomized studies. We assess modifications to existing approaches to assess bias, based on study design and methods of analysis.

What is the implication and what should change now?

- Current tools used to assess bias in systematic reviews can be modified to incorporate specific evaluation questions to assess nonrandomized studies with selection on unobservables. Work is underway to incorporate these approaches into Cochrane's risk of bias tool in nonrandomized studies of interventions.

inference being confounding and bias in selection of the reported result. In addition, the assessment of QEs requires greater qualitative appraisal of potential biases than RCTs, which in many cases may need to draw on advanced theoretical and statistical knowledge [4]. At the same time, QEs typically have a number of distinct advantages over experiments because they do not interfere in the natural data generation process [8].

Systematic critical appraisal, operationalized through ''risk of bias'' assessment, gives assurance of the credibility of the point estimates provided causal studies [9] and their trustworthiness for decision making [10]. Risk of bias tools provide transparency about the judgments made by reviewers when performing assessments. They are usually organized around particular domains of bias and provide specific ''signaling questions'' which enable reviewers to evaluate the likelihood of bias.

This paper discusses how to operationalize risk of bias assessment for QEs with selection on unobservables. A glossary of technical terms used is provided in the Appendix at www.jclinepi.com. Section 2 discusses internal validity, and Section 3 reviews existing risk of bias tools. Section 4 presents proposed evaluation criteria. Section 5 proposes an agenda for research in the further development of a risk of bias tool. Section 6 concludes.

2. Internal validity of QEs

Habicht et al. [11] distinguish probability evaluation designs, which are able to quantify with statistical precision the change in outcomes attributed to a treatment, from plausibility designs, which attempt to rule out observable confounding through use of a comparison group but are unable to address important sources of bias, in particular those arising from unobservables. These authors explicitly limit probability evaluations to RCTs. However, evidence is emerging which suggests QEs which use credible methods to address unobservable confounding can produce the same effect sizes as RCTs in pooled analysis (Table 1). We note that authors and journal editors may have incentives for selective publishing of favorable comparisons between randomized and nonrandomized studies. The examples presented in Table 1 are from systematic reviews (SRs) of socioeconomic interventions in low- and middle-income countries supported by the Campbell Collaboration International Development Coordinating Group (IDCG). The findings on experimental and quasi-experimental approaches are representative of the body of evidence in SRs supported by the IDCG. Other examples of comparisons of RCTs and QEs include Lipsey and Wilson [15] who provide a meta-analysis of North American social programs and Vist et al. [16] who compare RCTs and cohort studies in health care studies. Evidence is also available from (within-study) design replication—that is, studies which attempt to compare the same experimental treatment groups with nonrandomized comparison groups using quasi-experimental methods. One meta-study suggested significant differences between results from RCTs and QEs for US and European labor market programs [17]. However, design replications using well-conducted quasi-experimental methods, in which participation has been carefully modeled, have also shown the same results as the RCTs they are replicating [18,19].

As noted by Duvendack et al. [20], effect sizes estimated from nonrandomized studies may differ empirically from those from RCTs due to differences in the population sampled and the type of treatment effect estimated.

QEs modeling selection on unobservables account for confounding by design, either through knowledge about the method of allocation or in the methods of analysis used. They are considered more credible in theory than approaches based on unconfoundedness which rely solely on observable covariate adjustment [21,5,3].

In QE designs that use information about the allocation process to estimate a treatment effect, the ability of the

Table 1. Pooled effects of RCTs and credible quasi-experiments

Treatment	Design	Pooled odds ratio[a] (95% confidence interval)	$P > \|z\|$	Tau-sq	I-Sq (%)	Num obs
Cash transfer (vs. control)[b]	RCT	1.40 (1.21–1.61)	0.000	0.06	90.0	15
	QE	1.38 (1.25–1.52)	0.000	0.04	87.2	27
Education intervention (vs. standard intervention)[c]	RCT	1.33 (1.20–1.46)	0.000	0.02	90.9	43
	QE	1.34 (1.20–1.52)	0.000	0.02	96.7	16
Microcredit (vs. control)[d]	RCT	0.99 (0.93–1.05)	0.437	0.00	0.0	4
	QE	0.99 (0.88–1.12)	0.740	0.00	61.6	3

Abbreviations: RCT, randomized controlled trial; QE, quasi-experiment.

[a] Pooled odds ratios estimated by inverse-variance weighted random-effects meta-analysis. QEs included in the analyses use difference in differences, instrumental variables estimation, propensity score matching and regression discontinuity design.

[b] Baird et al. [12]; outcome is school enrollment.

[c] Petrosino et al. [13]; outcomes are school enrollment and attendance.

[d] Vaessen et al. [14]; outcome is "woman makes household spending decisions."

Source: author calculations based on reported data.

study to identify a causal relationship rests on assumptions that the variables which determine assignment are highly correlated with treatment status but not caused by the outcomes of interest nor confounded by any of the other causes of the change in outcome—that is, they are "exogenous." This is the same rationale on which randomized assignment is based; hence, we adopt the term "as-if randomized" [22] for quasi-experimental designs which are, in theory, able to account for all sources of confounding, including unobservables. We differentiate these designs from "nonrandomized" QEs which are only able to account for observable confounding and unobservables under particular conditions (Fig. 1).

"As-if randomized" QEs include instrumental variables, natural experiments, and discontinuity designs. In natural experiments, treatment may be assigned randomly due to "the forces of nature" or policy processes - for example, the Vietnam war draft lottery has been used to evaluate the effects of war on outcomes for conscripts [23,21]. Treatment may also be assigned quasi-randomly due to decisions in implementation or take-up, for example, where an arbitrary boundary determines service provision jurisdiction [24] or treatment practice [25]. Likewise quasi-randomness may occur due to errors in implementation [26]; see also [1]. Note that our definition of NE is narrower than Craig et al. [27] who apply the term to broader QE approaches [28].

In the special case of instrumental variables [29] and related approaches (e.g., "switching regression" models [30], researchers specify exogenous variables to model treatment decisions in multiple-stage regression (e.g., two-stage least squares or simultaneous equations maximum likelihood). Exogenous variables used in IV estimation include "randomized encouragement" [1], differences in implementation across groups [31] and, frequently, geographical factors such as distance, weather or climate conditions [32], and topography [33] (see [21] for a comprehensive overview of instrumental variables approaches). Instrumental variables methods are also used to analyze experimental data, for example, to account for noncompliance [34]; see [35] for an illustration of the approach in epidemiology. In the case of randomized encouragement, participants are exposed randomly by researchers to information about an intervention which itself is universally available but take-up is limited. Estimation is done using instrumental variables methods where the

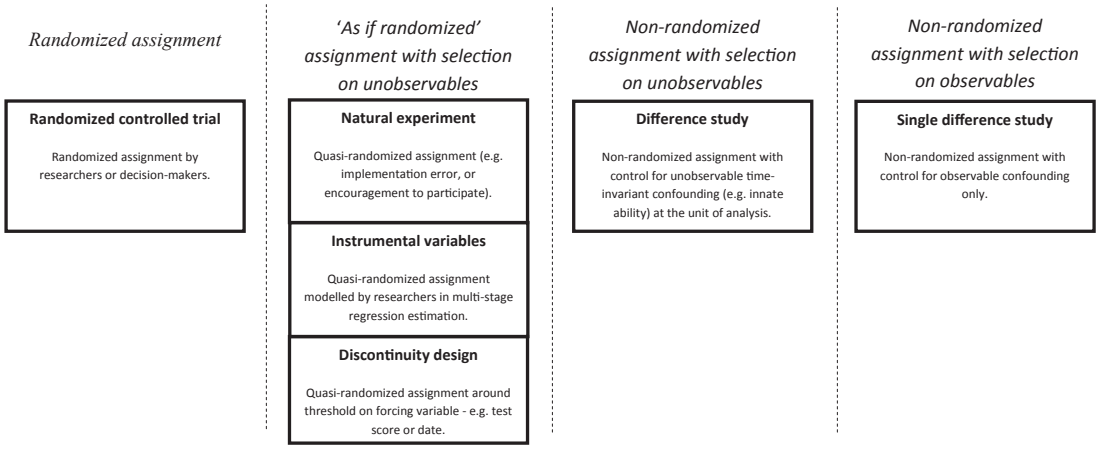

Fig. 1. Study design decision flow for studies of effects using statistical methods.

relationship of study interest is not the pragmatic question about the effect of such encouragement, but rather the mechanistic question about the effect of the intervention in people who are responsive to the encouragement.

"As-if randomized" QEs also include discontinuity studies which exploit local variation around a cutoff on an ordinal or continuous "forcing" variable used by decision makers to determine treatment. In the case of RDD treatment may be assigned by diagnostic test score [36], age [37] or even date of implementation, for example. The discontinuity is used to delimit treated and untreated observations as different units from the sample. In the case of ITS, where treatment is also determined date of implementation [38], untreated and treated observations are taken for the same sample units measured before and after implementation. ITS here refers to longitudinal panel data sets measured at the disaggregate level (i.e., the same medical cases or people measured multiple times before and after treatment [39]).

Where allocation rules are not exogenous, confounding must be controlled directly in adjusted statistical analyses. Methods such as DID (also called double differences), triple differences, and fixed-effects analysis of individual level longitudinal panel data—or pseudo-panels/repeated cross-section data under particular conditions [40]—enable adjustment for time-invariant unobservable confounding at the level of the unit of analysis by design [41]. However, these methods cannot control for time-varying unobservables even in theory. In contrast, single difference estimation applied to case–control, cohort, cross-sectional data, or PSM with matching on baseline characteristics, is not able in theory to control for time-varying or time-invariant unobservables, except in the special case of unconfoundedness (selection on observables). (The authors follow [3] in defining single difference studies as QEs.)

In Figure 1, we group these study designs and methods of analysis into three categories ordered from left to right according to a priori internal validity in addressing confounding. Randomized experiments and "as-if randomized" QEs are considered the most credible approaches in theory as they allow for selection on unobservables. Nonrandomized QEs with selection on unobservables are considered in theory more credible than nonrandomized studies relying solely on analysis of observables.

The choice of design for a comparative assessment should capture the information needed to classify a study in this proposed hierarchy (see also [1]). However, the extent to which designs produce valid causal inferences in practice also depends on the quality of implementation of the approach and the statistical conclusions drawn [42]. Particular flaws in implementation can lead to studies being assessed as being of lower quality than suggested by the a priori categories in Figure 1; indeed, we would expect many studies to be downgraded because the assumptions underlying the design are not met. Conversely, strong implementation might in rare cases lead to studies being assessed as of higher quality. An example would be an SD study where selection of participants is based on observable characteristics which are measured at baseline and appropriately modeled in the analysis.

Risk of bias assessments should therefore incorporate questions about design assessment and implementation of analysis [43]. Although the underlying domains of bias (e.g., confounding, sample selection bias, bias due to measurement error, and reporting bias, as defined below) are relevant across designs, the criteria used to verify them may differ based on the assumptions underlying the design. For instance, let Z be a variable determining assignment, T be a dummy variable representing treatment assignment, and Y be the outcome of interest. For randomized (RCTs) and "as-if randomized" (NE, IV and RD) studies, validity assessment should incorporate the following assumptions: predictable relationship between Z and T—in particular, nonzero and monotonic causal relationship between Z and T ("monotonicity") [44]; the relationship between treated units—Z for one treatment unit does not affect T for another treatment unit (crossovers), T for one treatment unit does not affect Y for another treatment unit (spillovers), and there is no variation in T across treatment units (e.g., due to measurement errors)—collectively referred to by econometricians as the "stable unit treatment value assumption" (SUTVA) [45]; and the relationship between Z and Y—that is, Z is not affected by Y or any of its causes and only affects Y through T (exogeneity, also called the "exclusion restriction" in instrumental variables literature). However, the "signalling questions" on which these propositions can be assessed will differ by study design. For example, the appropriate bandwidth around the cutoff in RDD, the model of autocorrelation in ITS, and the identification of observations in the region of common support in DID [46] (Fig. 2).

3. Review of critical appraisal tools

Many tools exist to facilitate risk of bias assessment of nonrandomized studies. Drawing on the systematic review by Deeks et al. [48] and a search of more recent literature, we selected and appraised relevant risk of bias tools according to the extent to which they identified evaluation criteria and signaling questions for QEs with selection on unobservables (Table 2). We included tools aiming to assess both randomized and nonrandomized studies [50,51,53,55,56,58,59,61] (the EPOC tool was developed drawing on the Cochrane risk of bias tool [62]). We also included tools aiming to appraise only nonrandomized studies [49,52,54,57,60].

Our assessment indicated that existing tools contain evaluation criteria for domains of bias that are relevant to QEs with selection on unobservables. However, most of the tools were not designed to assess causal validity of these studies, meaning that the "signaling questions" on which biases are evaluated were not sufficiently targeted,

Natural experiments and instrumental variables estimation:

- Correlation of assignment and treatment: assignment variable (instruments) and treatment status are monotonically correlated with sufficient variation in assignment variable (instruments).
- Confounding: assignment variable does not affect the outcome except by affecting treatment (the 'exclusion restriction').
- Independence of treatment and comparison observations across clusters (SUTVA is satisfied).
- Estimation: appropriate regression specification used for instrumental variables.

Regression discontinuity:

- Correlation of assignment and treatment: Forcing variable is continuous (or at least ordinal with sufficient values) and correlated with treatment at threshold.
- Confounding: Forcing variable is not confounded by other causes of the outcome (e.g., it is not used to determine allocation to another relevant intervention which affects outcome; threshold is not anticipated or manipulable by participants).
- Independence of treatment and comparison observations across clusters (SUTVA is satisfied).
- Estimation: Appropriate bandwidth around forcing threshold and regression specification.

Interrupted time series:

- Correlation of assignment and treatment: Intervention diffuses in population immediately or any process of anticipated or delayed diffusion is known and modeled.
- Confounding: Effects are not caused by other factors such as simultaneous implementation of other intervention; any time lag delaying effects is known and modeled.
- For controlled interrupted time series: Independence of treatment and comparison observations across clusters (SUTVA is satisfied).
- Estimation: Sufficient observations exist to estimate trends before and after, while ruling out other sources of variability such as autocorrelation (e.g. due to seasonality) or error structures can be credibly modeled.

Difference studies:

- Confounding: Differencing (or use of panel data regression) controls for observable and unobservable time-invariant confounding at the level of the unit of analysis. There is no unobservable time varying confounding differentially affecting outcome trajectories in treatment and comparison (the equal trends assumption).
- Independence of treatment and comparison observations across clusters (SUTVA is satisfied).
- Estimation: Comparable observations are used across groups (common support) or observations weighted accordingly.

Fig. 2. Assumptions underpinning internal validity of quasi-experiments with selection on unobservables. Sources: [41,3,47].

particularly in the domains of confounding and reporting biases. For example, randomization (sequence generation and allocation concealment) is usually the only method proposed to account for unobservable confounding. No single tool fully evaluated the internal or statistical conclusion validity of quasi-experimental designs, including the recent tool by Sterne et al. [57] which was operationalized for cohort studies. Only one tool addressed instrumental variables designs and statistical matching methods [53], and three tools presented signaling questions for discontinuity designs, of which the most comprehensive was [47]. Furthermore, most tools which addressed controlled before and after data (e.g., EPOC, n.d.) did not assess the degree to which time-varying unobservables at the level of the unit of analysis (e.g., patient, practitioner or health facility) were controlled using DID methods applied to sufficiently disaggregated data. Most tools that aimed to assess experimental and quasi-experimental studies did not enable consistent classification of experimental and quasi-experimental studies, or of different quasi-experimental designs, across similar evaluation criteria (e.g. [55]).

To take a recent example, the Cochrane Collaboration has developed a tool to assess risk of bias in nonrandomized studies of interventions [63]. That tool uses sensible evaluation criteria to assess risk of bias, with items grouped at the preintervention stage (baseline confounding and sample selection bias), during intervention (bias in measurement of interventions, e.g., due to problems of implementation fidelity or in recalling treatment status), and after the intervention has started (time-varying confounding, bias due to departures from intended interventions, bias due to missing data, bias in measurement of outcomes, and bias in selection of the reported result). But signaling questions to assess confounding focus on methods of observable covariate adjustment and are not sufficiently developed to account for selection on unobservables. Some sources of bias may not be relevant (e.g.,

Table 2. Assessment of experiments and quasi-experiments in existing critical appraisal tools

Tool	Experiment (RCT)	Natural experiment (NE)	Discontinuity design (RDD)	Interrupted time series (ITS)	Instrumental variables (IV)	Difference study (DID)
Cowley [49]	NA	N	N	P	N	N
EPOC [50]	Y	N	N	P	N	P[a]
Downs and Black [51]	Y	N	N	N	N	N
EPHPP [52]	Y	N	N	P	N	N
Hombrados and Waddington [53]	Y	P	P	N	Y	Y
Kim et al. [54]	NA	N	N	P	N	N
NICE [55]	Y	N	N	Y	N	N
Reisch et al. [56]	Y	N	N	N	N	N
Schochet et al. [47]	NA	NA	Y	P	NA	NA
Sterne et al. [57]	NA	P	N	P	P	P
Valentine and Cooper [58]	Y	N	P	P	N	P
SIGN 50 [59]	Y	N	N	N	N	N
Wells [60]	NA	N	N	P	N	N
West et al. [61]	Y	N	N	N	N	N

Abbreviations: Y, addresses study design and methods of analysis; P, partially addresses these; N, does not address; NA, not applicable.
[a] Includes controlled before and after only.

nonrandom attrition in a cross-sectional IV study). The tool instructs users to specify an unbiased "target randomized trial" [63] to which the particular nonrandomized study should be compared. This approach has been useful in getting reviewers from outside of the clinical trials community to think about sources of bias which they may previously have been unaware. However, there are instances where trials may be biased in ways which are not applicable to observational studies (e.g., performance bias due to Hawthorne effects, as discussed below).

To summarize, we are not aware of any single tool that sufficiently distinguishes control for unobservable confounding by design from control for observable confounding in analysis, for nonrandomized studies. Each tool addresses some of the potential biases for particular designs, but none provides the specific signaling questions needed to determine whether QEs are credible enough to recommend using the results in practice or policy. Application of these instruments is therefore likely to lead to inappropriate risk of bias assessment for QEs with selection on unobservables.

4. Evaluation criteria for QEs with selection on unobservables

In this section, we discuss evaluation criteria and potential signaling questions for quasi-experimental studies with selection on unobservables. Sterne et al. [63] categorize seven domains of bias which are relevant for nonrandomized studies: confounding, sample selection bias, bias due to missing data, bias in measurement of interventions, bias due to departure from intended interventions, bias in measurement of outcomes, and bias in selection of the reported result. These domains form the basis of evaluation criteria that can be used to operationalize risk of bias assessment for QEs. Our discussion focusses on how these domains apply to QEs, recognizing that the categories are also applicable for RCTs.

Confounding refers to the extent to which causality can be attributed to factors determining outcomes other than the intervention. Confounding factors that have been shown to influence outcomes include self-selection and program placement biases [63], as well as all nonintervention factors that affect the evolution of outcomes. Sources of confounding may be observable or unobservable, and time invariant or time varying. Studies using quasi-experimental approaches need to argue convincingly and present appropriate results of statistical verification tests, that the variable determining treatment assignment is exogenous to outcomes, and/or that the methods of analysis can control for unobservables [53]. For example, location is often endogenous since, at least in the long-term, people are able to move to gain access to better services. Hence, distance of participant to treatment facility may often not be a good instrumental variable [64]. Data permitting, it is also useful to make assessments of group equivalence at baseline according to observable covariates [65], under the assumption that these are correlated with unobservables, and to ensure common support is established [46]. Factors which may invalidate group equivalence during the process of implementation, such as time-varying confounding, should also be taken into account in estimation. For example, it is common for ITS to be applied to longitudinal datasets where the unit of analysis is clustered at aggregate levels of care (e.g., the health facility or district). In such cases, confounding by secular trends need to be assessed, for example with reference to a contemporaneous comparison group. In the case of DID, analysis can only adjust for time-invariant unobservable confounding at the unit of analysis. Hence, it is important to distinguish studies where data analysis is at the individual level, from those where data analysis is conducted at the aggregate level such as the community, hospital or higher.

Sample selection bias occurs where some eligible treatment units or follow-up periods are excluded from data collection or analysis, and this exclusion is correlated with outcome and intervention status. We define this effect of "selection into the study or analysis" differently from "selection bias" as usually defined in health care evaluation as

a special case of confounding (see above). Examples are nonrandom attrition and censoring of data (e.g., where outcomes data are not available due to mortality). Sterne et al. [57] refer to censoring as inception/lead time and immortal time biases; they are particularly important in retrospective studies and studies where baseline data are not available. Assessment is needed of the extent to which the design and methodology account for sample selection biases (e.g., through the use of Heckman error correction or the use of attrition-adjusted weights [66]). A related domain is bias due to missing data, which is a specific source of selection bias which also incorporates biases due to incomplete data collection (e.g., on outcomes, treatment status, or covariates measured at baseline and after) [63]. Biases due to differential attrition are potentially relevant only in prospective studies, but biases due to incomplete data collection are relevant for all designs.

Bias in measurement of interventions is not usually considered problematic where information is collected at the time of the intervention from sources not affected by the outcomes (e.g., enumerators). It is particularly problematic where information about treatment status is obtained after implementation from participants who may have an incentive to misreport or where recalling the intervention (e.g., its dose, frequency, intensity, or timing) is difficult [63]. This source of bias is most likely to occur where data are collected retrospectively.

Bias due to departures from intended interventions encompass crossovers, spillovers, and implementation fidelity. Crossovers or switches (including "contamination" of comparison groups) occur where individuals receive a treatment different from that assigned. Due to SUTVA, they are potentially problematic in all controlled studies including nonblinded prospective trials (and, e.g., double-blinded RCTs with an adaptive design where patients crossover if they do not improve sufficiently). Assessment should therefore be made of the extent to which these are accounted for in design or analysis. For example, in the case of RDD where the forcing variable does not precisely determine assignment (e.g. due to practitioner interference or anticipation effects by participants who have knowledge about the assignment mechanism [3]), the design is referred to as "fuzzy" (as opposed to "sharp") discontinuity design. Depending on the degree of noncompliance, estimation may done through intention-to-treat or instrumental variables. Anticipation effects may also complicate analysis in ITS, where knowledge of future intervention may change behaviour and outcomes prior to implementation (Julian Higgins, personal communication). Spillovers occur when members of the comparison group are exposed to treatment indirectly, through contact with treated individuals, and are potentially problematic for all controlled studies. Cluster-level analysis may be required to ameliorate these sources of bias and/or an assessment of the geographical or social separation of groups may be needed [53].

Bias in measurement of outcomes due to recall and courtesy biases is potentially problematic in all studies where outcomes data are self-reported. But other forms of motivational bias are only likely to arise in prospective studies. The classic case is the presence of Hawthorne and John Henry effects affecting motivation of participants when they are aware they are part of a trial (particularly when they are subjected to repeated measurement). As another example, "survey effects" may operate whereby groups are sensitized to information that affects outcomes through survey questions and then subjected to repeated measurement [67]. Such effects are less likely to affect motivation where data are collected outside of a trial situation with a clear link to an "intervention" and unlikely to be relevant when data are collected at one period of time as in a retrospective cross-sectional only [53]. Blinding is frequently advocated to reduce bias in outcomes measurement. Although it may be impossible to prevent participant knowledge of intervention status (especially in evaluations of health systems interventions), blinding of outcome assessors and data analysts usually is feasible, though often not undertaken. Blinding of participants may also be less relevant for socioeconomic and health systems interventions, where expectations (such as placebo effects) may form an important mechanistic component in the process of behavior change.

Bias in selection of the reported result corresponds to selective reporting of outcomes (e.g., among multiple possible outcomes collected), selective reporting of results from subgroups of participants, or selective reporting of methods of analysis (e.g., where multiple estimation strategies or specifications are used) [68,57]. These types of bias are particularly likely to be prevalent in retrospective studies based on observational data (e.g., with many IV analyses) but may also arise in prospective QEs where the method of analysis or outcomes is chosen based on results. Presence of a study protocol (preanalysis plan) can help determine the likelihood of bias, although it is recognized that many such studies still do not contain such plans nor is it usually possible to fully specify all statistical methods in advance. However, risk of bias assessment can also aim to incorporate the use of unusual or uncommon methods of analysis [53].

5. Operationalizing the approach

Further development of a tool or tools to assess QEs with selection on unobservables should, first, aim to build on the bias domains and signaling questions in existing tools used by reviewers, in particular those articulated by Sterne et al. [63]. Second, the tool should address both the conceptual and statistical assumptions underpinning validity. This means that appraisals of, for example, the likely exogeneity of quasi-randomization in the confounding domain will need to be incorporated. The evaluation of assumptions underpinning QEs is notoriously more difficult than that of RCTs, relying to a greater extent on what we might call "qualitative judgment" informed by both advanced statistical and substantive theoretical knowledge.

Appraisal by multiple reviewers and interrater reliability assessment are therefore crucial [9].

Third, an integrated assessment tool, covering multiple study designs, should incorporate both articulation of the study design [1]. Fourth, analysis should be based on what is being reported regarding the assumptions of the designs and the methods with which they are addressed [43]. However, the process should not operate to penalize studies with transparent reporting over and above studies with limited reporting (e.g. of statistical testing that might undermine the case for causal identification). Hence, the need for consistent assessment across studies and incorporation of appropriate responses to signalling questions (e.g., including both ''unclear'' and ''not reported'' response categories).

Finally, it is likely that some of the signaling questions used to operationalize evaluation of bias will be design specific, in particular for confounding and reporting domains. For natural experiments and instrumental variables, this will require qualitative appraisal of the exogeneity of the identifying variable or instrument. For instrumental variables, the assessment should also incorporate the significance or goodness-of-fit of the first-stage instrumenting equation, the individual significance of the instruments and results of an overidentifying test if applicable [53]. For regression discontinuity, the assessment should incorporate whether the forcing variable is continuous or at least ordinal with sufficient values [47], the degree to which assignment is exogenous (i.e., not manipulable by participants in response to incentives), comparison of covariate means either side of the threshold, and an assessment of appropriate specification (bandwidth and use of weighting for matches further from the assignment threshold) and functional form (e.g., step vs. slope, linear or nonlinear relationship between forcing variable and outcome). For ITS, the assessment should incorporate functional form and bias due to confounding, which may be done with respect to a control group or a ''placebo outcome'' which is not expected to be affected by the intervention (see [1]). For DID, assessments are needed of the unit of analysis at which the differencing occurs (determining whether time-invariant unobservables are controlled at, e.g., patient, practitioner, health facility, or higher level), and the existence of equal trends in outcomes before intervention across treatment and comparison groups (an indicator of whether unobservable confounders are changing differentially across group) [41].

As in the case of randomized studies, information needed to inform risk of bias judgments in QEs must be collected from the studies (see [9]; p. 194-197; [57]). When specific information is unknown, reviewers may attempt to obtain such information from the primary study authors. Some reviewers may believe that absence of such information is enough to exclude a study from a review. This should be explicitly stated as part of the inclusion criteria. Where studies are eligible for inclusion by stated design alone, the presence or absence of this information should be incorporated into the risk of bias assessment, as noted above, and methods such as meta-regression can be used to explore systematic differences between primary studies that do or do not report information.

The information obtained from risk of bias instruments can be used in a variety of ways [69]. Ahn and Becker [70] and Herbison et al. [71] present evidence that meta-analysis should not be weighted by quality scores. Determining overall risk of bias across categories is complicated, because the degree of bias is a latent construct (i.e., one that is not directly observable or measureable), but can be useful [72]. Although evidence suggests it is not appropriate to determine overall bias using weighted quality scales [73], reviewers have shown that it is possible to assess overall bias based on transparent decision criteria (e.g., the SRs reported in Table 1). Others prefer to code separate indicators of particular biases to serve as potential moderator variables.

6. Conclusions

Current tools used by reviewers do not provide the means to evaluate consistently and appropriately the credibility of quasi-experimental studies with selection on unobservables. The paper justifies the further development of a comprehensive tool for nonrandomized studies [63] and suggests how it might incorporate QEs. The tool should be operationalized to recognize explicitly credible quasi-experimental approaches like difference studies, instrumental variables, interrupted time series, natural experiments, and regression discontinuity designs. It should assess these using consistent evaluation criteria across bias domains which incorporate both the assumptions of the study design and the implementation of the approach. It is likely that different signaling questions will be required for different designs, particularly to address confounding and the reporting of appropriate statistical analyses.

Acknowledgment

The authors would like to acknowledge the contribution of NIHR Biomedical Research Unit in Cardiovascular Disease. Thanks are due to participants at the Alliance for Health Systems Research workshop on quasi-experimental studies at Harvard School of Public Health, November 2013, the Campbell Collaboration Methods Group Symposium in Belfast, May 2014, and the Cochrane Methods workshop in Hyderabad, September 2015. We also thank Ian Shrier for helpful comments.

Supplementary data

Supplementary data related to this article can be found at http://dx.doi.org/10.1016/j.jclinepi.2017.02.015.

References

[1] Reeves BC, Wells GA, Waddington H. Quasi-experimental study designs series-paper 5: a checklist for classifying studies evaluating the effects on health interventions-a taxonomy without labels. J Clin Epidemiol 2017;89:30−42.

[2] Rockers PC, Rottingen JA, Shemilt I, Tugwell P, Bärnighausen T. Inclusion of quasi-experimental studies in systematic reviews of health systems research. Health Policy 2015;119(4):511−21.

[3] Shadish W, Cook T, Campbell D. Experimental and quasi-experimental designs for generalized causal inference. Belmont, CA: BROOKS/COLE CENGAGE Learning; 2002.

[4] 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.

[5] Imbens GM, Wooldridge JM. Recent developments in the econometrics of program evaluation. J Econ Lit 2009;47(1):5−86.

[6] Higgins JPT, Ramsay C, Reeves BC, Deeks JJ, Shea B, Valentine JC, 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 2012;4(1): 12−25.

[7] Rubin D. Estimating causal effects of treatments in randomized and nonrandomized studies. J Educ Psychol 1974;66(5):689.

[8] 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.

[9] Higgins JPT, Green S. Cochrane handbook for systematic reviews of interventions, Version 5.0.0. London: John Wiley and Sons; 2011.

[10] Chalmers I. The development of fair tests of treatments. Lancet 2014;383:1713−4.

[11] Habicht J-P, Victora CG, Vaughan JP. Evaluation designs for adequacy, plausibility and probability of public health programme performance and impact. Int J Epidemiol 1999;28:10−8.

[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;9(8):1−124.

[13] Petrosino A, Morgan C, Fronius TA, Tanner-Smith EE, Boruch RF. Interventions in developing nations for improving primary and secondary school enrollment of children: a systematic review. Campbell Syst Rev 2012;8(19):1−192.

[14] Vaessen J, Rivas A, Duvendack M, Palmer-Jones R, Leeuw FL, Van Gils G, et al. The effects of microcredit on Women's control over household spending in developing countries: a systematic review and meta-analysis. Campbell Syst Rev 2014;10(8):1−205.

[15] Lipsey MW, Wilson DB. The efficacy of psychological, educational, and behavioural treatment: confirmation from meta-analysis. Am Psychol 1993;48:1181−209.

[16] Vist GE, Bryant D, Somerville L, Birminghem T, Oxman AD. Outcomes of patients who participate in randomized controlled trials compared to similar patients receiving similar interventions who do not participate. Cochrane Database Syst Rev 2008; 1−106.

[17] Glazerman S, Levy D, Myers D. Nonexperimental versus experimental estimates of earnings impacts, Annals of the Academy of Political and Social Sciences 2003. Available at http://www.povertyactionlab.org/doc/non-experimental-vs-experimental-estimates. Accessed December 3, 2015.

[18] Cook T, Shadish W, Wong V. Three conditions under which experiments and observational studies produce comparable causal estimates: new findings from within-study comparisons. J Policy Anal Manage 2008;27(4):724−50.

[19] Hansen H, Klejntrup N, Andersen O. A comparison of model-based and design-based impact evaluations of interventions in developing countries, FOI Working Paper 2011/16 2011. Available at http://okonomi.foi.dk/workingpapers/WPpdf/WP2011/WP_2011_16_model_vs_design.pdf. Accessed April 14, 2017.

[20] Duvendack M, Garcia Hombrados J, Palmer-Jones R, Waddington H. Assessing 'what works' in international development: meta-analysis for sophisticated dummies. J Develop Effect 2012;4(3):456−71.

[21] Dunning T. Natural experiments in the social sciences: a design-based approach. Cambridge: Cambridge University Press; 2012.

[22] Dunning T. Design-based inference: beyond the pitfalls of regression analysis?. In: Collier D, Brady H, editors. Rethinking social inquiry: diverse tools, shared standards. 2nd Edition. Lantham, MD: Rowman and Littlefield; 2010.

[23] Angrist J, Pischke S. Mostly Harmless Econometrics: An empiricist's companion. New Jersey: Princeton University Press; 2009.

[24] Snow J, Richardson BW. Snow on cholera: being a reprint of two papers by John Snow, MD, together with a biographical memoir by B.W. Richardson, and an introduction by Wade Hampton Frost, MD. New York: Hafner; 1965.

[25] Zafar SN, Libuit L, Hashmi ZG, Hughes K, Greene WR, Cornwell EE III, et al. The sleepy surgeon: does night time surgery for trauma affect mortality outcomes? Am J Surg 2015;209:633−9.

[26] Morris S, Olinto P, Flores R, Nilson E, Figueiró A. 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(9): 2336−41.

[27] Craig P, Cooper C, Gunnell D, Haw S, Lawson K, Macintyre S, et al. Using natural experiments to evaluate population health interventions: guidance for producers and users of evidence [online]. London: Medical Research Council; 2011. Available at https://www.mrc.ac.uk/documents/pdf/natural-experiments-to-evaluate-population-health-interventions/. Accessed November 4, 2015.

[28] King G. Public policy for the poor? A randomised assessment of the Mexican universal health insurance programme. Lancet 2009;373: 1447−54.

[29] Zohoori N, Savitz DA. Econometric approaches to epidemiologic data: relating endogeneity and unobserved heterogeneity to confounding. Ann Epidemiol 1997;7:251−7.

[30] Lockshin M, Sajaia Z. Maximum likelihood estimation of endogenous switching regression models. Stata J 2004;3:282−9.

[31] Wang H, Norton EC, Rozier RG. Effects of the state children's health insurance program on access to dental care and use of dental services. Health Serv Res 2007;42:1544−63.

[32] Lawlor DA, Davey Smith G, Mitchell R, Ebrahim S. Adult blood pressure and climate conditions in infancy: a test of the hypothesis that dehydration in infancy is associated with higher adult blood pressure. Am J Epidemiol 2006;163:608.

[33] Duflo E, Pande R. Dams. Q J Econ 2007;122(2):601−46.

[34] Imbens GW, Angrist JD. Identification and estimation of local average treatment effects. Econometrica 1994;62(2):467−75.

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

[36] Bor J, Moscoe E, Mutevedzi P, Newell ML, Bärnighausen T. Regression discontinuity designs in epidemiology: causal inference without randomized trials. Epidemiology 2014;25:729−37.

[37] Card D, Dobkin C, Maestas N. Does medicare save lives? Q J Econ 2009;124(2):597−636.

[38] Everitt DE, Soumerai SB, Avorn J, Klapholz H, Wessels M. Changing surgical antimicrobial prophylaxis practices through education targeted at senior department leaders. Infect Control Hosp Epidemiol 1990;11(11):578−83.

[39] Lopez Bernal JA, Gasparrini A, Artundo CM, McKee M. The effect of the late 2000s financial crisis on suicides in Spain: an interrupted time-series analysis. Eur J Public Health 2013;23:732−6.

[40] Verbeek M. Pseudo-panels and repeated cross-sections. In: Matyas L, Sevestre P, editors. The econometrics of panel data. Berlin Heidelberg: Springer-Verlag; 2008.

[41] Gertler P, Martinez S, Premand P, Rawlings L, Vermeersch C. Impact evaluation in practice. Washington DC: World Bank; 2016.
[42] Vandenbroucke JP. Is there a hierarchy of methods in clinical research? J Mol Med 1989;67(10):515−7.
[43] Littell J, Corcoran J, Pillai V. Systematic reviews and meta-analysis. New York: Oxford University Press; 2008.
[44] Bound J, Jaeger DA, Baker R. Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak. J Am Stat Assoc 1995; 90:443−50.
[45] Chiba Y. Bias analysis of the instrumental variable estimator as an estimator of the average causal effect. Contemp Clin Trials 2010; 31:12−7.
[46] Heckman J. Characterizing selection bias using experimental data. Econometrica 1998;66(5):1017−98.
[47] Schochet P, Cook T, Deke J, Imbens G, Lockwood JR, Porter J, et al. Standards for regression discontinuity designs. Princeton, NJ: Mathematica Policy Research Report; 2010.
[48] Deeks J, Dinnes R, D'Amico R, Sowden AJ, Sakarovitch C, Song F, et al. Evaluating non-randomised intervention studies. Health Technol Assess 2003;7. iii−x, 1−173.
[49] Cowley DE. Protheses for primary total hip replacement: a critical appraisal of the literature. Int J Technol Assess Health Care 1995; 11(4):770−8.
[50] Cochrane Effective Practice and Organisation of Care Group (EPOC), undated. Suggested risk of bias criteria for EPOC reviews. Available at https://epoc.cochrane.org/sites/epoc.cochrane.org/files/public/uploads/Suggested%20risk%20of%20bias%20criteria%20for%20EPOC%20reviews.pdf. Accessed April 14, 2017.
[51] Downs S, Black N. The feasibility of creating a checklist for the assessment of the methodological quality both of randomised and non-randomised studies of health care interventions. J Epidemiol Community Health 1998;52:377−84.
[52] Effective Public Health Practice Project (EPHPP), undated. Quality assessment tool for quantitative studies. Available at http://www.ephpp.ca/PDF/Quality%20Assessment%20Tool_2010_2.pdf. Accessed April 14, 2017.
[53] Hombrados JG, Waddington H. A tool to assess risk of bias for experiments and quasi-experiments in development research. New Delhi: Mimeo. The International Initiative for Impact Evaluation; 2012.
[54] Kim SY, Park JE, Lee YJ, Seo H-J, Sheen S-S, Hahn S, et al. Testing a tool for assessing the risk of bias for nonrandomized studies showed moderate reliability and promising validity. J Clin Epidemiol 2013;66:408−14.
[55] National Institute for Health and Clinical Excellence (NICE). Quality appraisal checklist − quantitative intervention studies. In: Methods for the development of NICE public health guidance (second edition), April 2009, NICE, London.
[56] Reisch J, Tyson J, Mize S. Aid to the evaluation of therapeutic studies. Pediatrics 1989;84(5):815−27.
[57] Sterne JAC, Higgins JPT, Reeves BC, on behalf of the development group for ACROBAT-NRSI. A Cochrane risk of bias assessment tool: for non-randomised studies of interventions (ACROBAT-NRSI). Version 1.0.0 2014. Available at http://www.riskofbias.info. Accessed September 24, 2014.
[58] Valentine J, Cooper H. A systematic and transparent approach for assessing the methodological quality of intervention effectiveness research: the Study Design and Implementation Assessment Device,. Psychol Methods 2008;13(2):130−49.
[59] SIGN. SIGN 50: A guideline developer's handbook. Revised Edition. Edinburgh: Scottish Intercollegiate Guidelines Network; 2011. Available at http://www.sign.ac.uk/pdf/sign50nov2011.pdf. Accessed December 1, 2015.
[60] Wells G, Shea B, Connell DO', Peterson J, Welch V, Losos M, et al., Undated. The Newcastle-Ottawa Scale (NOS) for assessing the quality of nonrandomised studies in meta-analyses. Available at http://www.ohri.ca/programs/clinical_epidemiology/oxford.asp. Accessed April 14, 2017.
[61] West S, King V, Carey TS, Lohr KN, McKoy N, Sutton SF, et al. Systems to rate of strength of scientific evidence. Evidence Report/-Technology Assessment Number 47 2002: AHRQ Publication No. 02−E016.
[62] Higgins JPT, Altman DG, Gøtzsche PC, Jüni P, Moher D, Oxman AD, et al. The Cochrane Collaboration's tool for assessing risk of bias in randomised trials. BMJ 2011;343:d5928.
[63] Sterne JAC, Hernan M, Reeves BC, Savović J, Berkman ND, Viswanathan M, et al. ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions. BMJ 2016;355:i4919.
[64] Newhouse JP, McClellan M. Econometrics in outcomes research: the use of instrumental variables. Annu Rev Public Health 1998; 19:17−34.
[65] Hansen BB. Covariate balance in simple, stratified and clustered comparative studies. Stat Sci 2008;23(2):219−36.
[66] Fitzgerald J, Gottschalk P, Moffitt R. An analysis of sample attrition in panel data: the Michigan panel study of income dynamics. J Hum Resour 1998;33(2):251−99.
[67] Zwane AP, Zinman J, van Dusen E, Pariente W, Null C, Miguel E, et al. Being surveyed can change later behavior and related parameter estimates. Proc Natl Acad Sci U S A 2011;108:1821−6.
[68] Rothstein H, Sutton A, Borenstein M, editors. Publication bias in meta-analysis: prevention, assessment and adjustments. London: Wiley; 2005.
[69] Ioannidis J. Meta-research: the art of getting it wrong. Res Synth Methods 2011;1(3−4):169−84.
[70] Ahn S, Becker BJ. Incorporating quality scores in meta-analysis. J Educ Behav Stat 2011;36(5):553−85.
[71] Herbison P, Hay-Smith J, Gillespie WJ. Adjustment of meta-analysis on the basis of quality scores should be abandoned. J Clin Epidemiol 2006;59:1249−56.
[72] Guyatt GH, Oxman AD, Akl EA, Kunz R, Vist G, Brozek J, et al. GRADE guidelines: 1. Introduction—GRADE evidence profiles and summary of findings tables. J Clin Epidemiol 2011;64:383−94.
[73] Jüni P, Witschi A, Bloch R, Egger M. The hazards of scoring the quality of clinical trials for meta-analysis. JAMA 1999;282: 1054−60.