JECH Online First, published on August 6, 2010 as 10.1136/jech.2008.082602

Alternatives to randomisation in the evaluation of public health interventions: design challenges and solutions

C P Bonell, J Hargreaves, S Cousens, D Ross, R Hayes, M Petticrew, B R Kirkwood

London School of Hygiene and Tropical Medicine, London, UK

Correspondence to
Chris Bonell, Public and Environmental Health Research Unit, London School of Hygiene and Tropical Medicine, Keppel Street, London WC1E 7HT, UK; chris.bonell@lshtm.ac.uk

Accepted 7 January 2009

ABSTRACT

Background There has been a recent increase in interest in alternatives to randomisation in the evaluation of public health interventions. We aim to describe specific scenarios in which randomised trials may not be possible and describe, exemplify and assess alternative strategies.

Methods Non-systematic exploratory review.

Results In many scenarios barriers are surmountable so that randomised trials (including stepped-wedge and crossover trials) are possible. It is possible to rank alternative designs but context will also determine which choices are preferable. Evidence from non-randomised designs is more convincing when confounders are well-understood, measured and controlled; there is evidence for causal pathways linking intervention and outcomes and/or against other pathways explaining outcomes; and effect sizes are large.

Conclusion Non-randomised trials might provide adequate evidence to inform decisions when interventions are demonstrably feasible and acceptable, and where evidence suggests there is little potential for harm, but caution that such designs may not provide adequate evidence when intervention feasibility or acceptability is doubtful, and where existing evidence suggests benefits may be marginal and/or harms possible.

EVALUATING THE EFFECTS OF PUBLIC HEALTH INTERVENTIONS: BARRIERS TO RANDOMISED TRIALS AND ALTERNATIVE OPTIONS

Randomised controlled trials (RCTs) are widely regarded as the 'gold standard' for estimating the causal effects of public health interventions on predefined outcomes in a defined population.[1 2] However, there has been increasing interest in non-randomised evaluations of public health interventions where RCTs are considered unfeasible[3–6] and guidelines for reporting these.[7 8] It has been argued that for certain decisions it is reasonable to infer intervention effects from non-randomised studies and that dismissing designs other than the RCT might lead to a marginalisation of intervention types not amenable to RCTs.[3–5] We organised a multi-disciplinary symposium in London in 2006 to review these issues and identify practical solutions. Our two papers summarise the arguments presented, drawing on examples from high- and low-income countries.

It is important to determine which categories of intervention are not amenable to RCTs in order that decisions to use evidence from non-randomised designs are made based not on a wholesale or haphazard 'lowering of the bar' regarding standards of evidence, but from a considered assessment as to what interventions require this. It is also important to assess in detail the various alternatives to RCTs in order to determine their relative merits in different scenarios. Existing reviews have not aimed to describe comprehensively the diversity of different alternative designs available, but instead have aimed to discuss overall levels of evidence[3 4] or focused on particular alternative designs.[6] We conclude by considering for what sorts of decisions non-randomised evaluations might provide adequate evidence and, conversely, when decisions require RCTs. This paper focuses on design while a companion paper considers analysis.

KEY FEATURES OF RCTS OF PUBLIC HEALTH INTERVENTIONS

RCTs are experiments that compare outcomes measured in prospective follow-up between those randomly allocated to receive an intervention and the control group randomly allocated to receive the comparison condition (generally currently accepted standard care). RCTs of public health interventions often allocate clusters of individuals (eg, villages or schools) to intervention or control. The strength of RCTs comes in their capacity to ensure a 'fair comparison' between intervention and control groups, which are 'balanced' and could be expected to experience comparable outcomes in the absence of intervention.

This paper will consider barriers and alternatives to three key features of RCTs: random allocation, control groups and prospective follow-up, only the first of which is unique to RCTs. RCTs also share with other evaluative designs additional features not discussed here. First, many RCTs hide treatment allocation from participants, providers and/or researchers to ensure measurement errors are non-differential by allocation. Although 'blinding' of providers and participants is sometimes impossible with public health interventions,[9] this is not discussed further because in such scenarios all that can be done is to ensure as far as possible that standardised measurement procedures are used in all groups, and that blinding is maintained—for example, among those managing and analysing data. Second, RCTs aim to include sufficient individuals or clusters of individuals. This should provide the statistical power to maximise the chances of detecting true effects and the statistical significance to minimise the chances of apparent effects arising by chance. Adequate

sample size is not discussed further because there are no obvious alternatives.

Barriers to key features of RCTs

Policy makers or providers may believe in the value of an intervention for certain individuals or groups, often regardless of its actual evidence base, and, therefore, oppose random allocation because it prevents their judging allocation. Similarly, clients may have preferences and oppose randomisation. This may arise particularly with interventions associated with ideological beliefs, such as some forms of psychotherapy and community development,[10 11] and might explain the non-randomised evaluation of the 'Healthy School and Drugs' project[12] (Box 1). Where there is demonstrable uncertainty about intervention benefits that is 'equipoise' regarding intervention effects, it may be possible to persuade providers that evaluation is required and, while they should decide who are suitable candidates, who within this pool receives the intervention will be determined randomly.[13] Client choice has been accommodated in 'preference trials', which enable clients to opt for random allocation or choose their preferred intervention, although it is debated which groups should then be compared in analysis.[14]

Random allocation may also be opposed because providers or policy makers are so keen to demonstrate the effectiveness of an intervention that they want it implemented in the most promising contexts while control units are not chosen for these reasons, introducing bias. Again where equipoise exists, evaluators might persuade intervention advocates that trial evidence should aid future advocacy.[15]

Randomisation might also be blocked where policy makers decide pilots must occur with the most needy areas or individuals.[2] In such scenarios, where providers assign interventions on the basis of a need score, intervention effect estimates might be derived from a 'regression-discontinuity' analysis that explores the shape of the association between the measure of need and the outcome of interest after introduction of the intervention. However, this approach makes several assumptions—notably that the shape of the association between the measures of need and outcome is known.[6] Finally, random allocation will also be impossible when decisions have already been made about

Box 1 Examples of non-randomised evaluations

Healthy School and Drugs project[12]
Intervention—(i) classroom drugs education, (ii) committee involving parents and teachers to coordinate drugs policy, (iii) new rules on substance use and (iv) support for pupils using drugs.
Barriers to RCT design—unclear, but perhaps potential participants were unwilling to undergo random allocation.
Evaluation design—non-randomised prospective concurrent control study with nine intervention and three control schools. Analysis adjusted for baseline differences in pupils' demographic factors (not reported which) and baseline substance-use knowledge and behaviour, but no account taken of clustering in analysis.
Outcomes—self-reported substance-use knowledge and behaviour
Results—some apparent effects, particularly on alcohol use. There may have been potential for unmeasured confounding; for example, from inter-school differences in academic achievement, attitude to school, institutional management etc.

Integrated Management of Childhood Illness (IMCI), Tanzania[16]
Intervention—guidelines, training and improved systems for IMCI 1997–2002.
Barriers to RCT design—decisions to implement IMCI made before study; only some clusters had surveillance systems.
Evaluation design—non-randomised comparison of data from routine mortality surveillance, household surveys and health facility surveys in two intervention and two control districts (total population ~1.2 million) with similar baseline child mortality rates; with process measures and checks for other potential influences (such as bed-net provision).
Outcome—health and survival of children aged less than 5 y.
Results—mortality rates lower in intervention clusters, but too few clusters to exclude chance as an explanation.

Childhood immunisation with pneumococcal conjugate vaccine, USA[19]
Intervention—national introduction of routine childhood immunisation with pneumococcal conjugate vaccine.
Barriers to RCT design—nationwide introduction of intervention for which vaccine efficacy for the prevention of invasive pneumococcal disease previously established.
Evaluation design—time-series analysis comparing post-introduction trends with expected trends (based on admissions prior to vaccine introduction). Trends in dehydration admissions also examined to explore alternative hypothesis that any apparent effects were merely the result of changes in the sampling of data on hospital admissions or of changes in healthcare coverage.
Outcome—monthly admissions for all-cause and pneumococcal pneumonia in the general population (using routine data from Nationwide Inpatient sample).
Results—relative decline in hospital admissions for relevant outcomes compared to predicted trends. No change in admissions for dehydration compared to predicted trends.

Mass-media family-planning intervention in Nepal[20]
Intervention—radio soap opera broadcast on the national radio, designed to promote the concept of a 'well planned family' and increase demand for family planning services.
Barriers to RCT design—national introduction of intervention.
Evaluation design—nationally representative, cross-sectional survey of ever-married women.
Exposure—woman recalls listening to the soap opera in the 6 months prior to interview.
Outcome—woman currently using a modern contraceptive.
Results—differed according to form of analysis (discussed further in companion paper).

where/to whom an intervention will be delivered—for example in the evaluation of the Integrated Management of Childhood Illness programme in Tanzania (see Box 1).[16]

In some cases, it may be impossible to have a control group, randomised or otherwise. This might arise where policy makers or practitioners believe an intervention is beneficial and no one in need should be denied it. Again, where equipoise genuinely exists, it should be ethical to undertake an RCT and persuade opponents of the value of this. However, grey areas exist: an intervention might be shown to be effective in one setting but uncertainty remains as to whether effects will translate to a new setting. This will depend on the complexity of the intervention and of the causal pathway from intervention to outcomes, dissimilarities in infrastructure and client characteristics[3 17] and, for infectious diseases, population differences in transmission dynamics.

Advocates of an intervention may find stepped-wedge or crossover trials more acceptable. Stepped-wedge RCTs stagger the introduction of an intervention, randomising the order of receipt.[18] In crossover RCTs, all participants receive the intervention for a period and the control condition for a period; randomisation determines the order. The latter is only useful in evaluating acute effects and in scenarios where it is acceptable to withdraw interventions after a period of delivery.

Another scenario is where intervention effects on primary outcomes are known but effects on other secondary but important outcomes are not. This rendered unethical any control-group study of routine childhood pneumococcal conjugate immunisation on pneumonia in the general population (Box 1) because effects on pneumonia incidence, but not on the population burden of disease, were known.[19] Where a social intervention's effects, for example on income[13] or legal rights,[11] are known but health effects are not, true equipoise may not exist, rendering control groups unethical.[13] Judgements here depend on the importance of the known benefits.[11]

Control groups will also generally be impossible where an intervention is already delivered as standard across an entire area,[9] such as with the Nepalese family planning intervention[20] (Box 1), since policy makers will usually be unwilling to withdraw it even where equipoise remains.[11] Where an intervention has yet to be delivered, control groups will also be impossible where it is legally, bureaucratically or practically necessary for delivery to be consistent across an entire state or nation—for example, with laws and regulations, welfare benefits or mass media.[13] The possibility of 'contamination' is also sometimes cited as a reason not to rely on control groups[21] although the effects of this can be reduced by employing a cluster design.[22]

Longitudinal follow-up of participants from pre-intervention baseline measures to post-intervention outcomes may be impossible when an evaluation begins only after an intervention has been delivered to a population or where policy makers or evaluation funders are reluctant to have lengthy periods of observation pre-implementation. Longitudinal follow-up may also be difficult when there are long gaps between intervention and manifestation of key outcomes, as is often the case with prevention[23] or where outcomes are rare.[21] Politicians may be uninterested in studies lasting longer than their probable period in office.

A final barrier to RCTs of public health interventions is lack of funding, particularly where trials aim to detect relatively small effects and require large samples. Sometimes it can be argued that only an RCT can adequately address a critical evidence gap. Our discussion section considers in what scenarios this is so. Funders may sometimes be prepared to fund smaller, cheaper 'non-inferiority' RCTs aiming to detect whether intervention benefits are equivalent to/better than current practice, although how best to analyse these is debated.[24]

ALTERNATIVES TO RANDOM ALLOCATION

When random allocation is not possible it may instead be possible to employ a control group with prospective matching and/or post hoc adjustment for potential confounders (table 1). The latter was used in the evaluation of the Dutch 'Healthy School and Drugs' intervention where a number of potential confounders such as attitude to smoking were examined.

The disadvantage of these options is they cannot control for unmeasured or imperfectly measured confounders. Among other threats to internal validity, the 'Healthy School and Drugs' project did not adjust for pupils' attitudes to school despite evidence that this might be a confounder.[25] Inadequate reporting of how potential confounders are identified has previously been identified as a deficit in many epidemiological studies[7] and applies equally to evaluations. Comprehensive matching/adjustment on all potentially important confounders is likely to be difficult when evaluations rely on routine data from intervention and/or control groups and when it is necessary to make adjustment for cluster-level variables but only a small number of clusters have been enrolled, as was the case with the Integrated Management of Childhood Illness evaluation. Confounding can lead to underestimates of effects[26] (for example, when intervention recipients' greater needs are not sufficiently considered) or, probably more commonly,[27–29] overestimates (for example, where intervention recipients' lesser needs or greater uptake of the intervention are insufficiently considered). For example, non-random studies of the association between vitamin-A deficiency and mother-to-child HIV-transmission reported associations, but RCTs of vitamin-A supplementation found no evidence of effect.[30] Our second paper considers other recently proposed analytic strategies to minimise confounding.

ALTERNATIVES TO CONTROL GROUPS

When it is not possible to recruit a prospective control group, it may still be possible to compare outcomes in a study population with rates in the general population. Where rates change in the intervention population but not the general population, this provides some evidence for intervention effects although confounding and regression to the mean (see below) may introduce bias. This approach has been used in studies evaluating the effects of new roads on mortality and injuries.[31]

In the absence of external comparison, 'before–after studies' may be possible. Such studies are vulnerable to confounding from secular and maturational trends as well as contemporaneous influential events. The extent to which these undermine an evaluation depends on context: secular trends are less problematic where rates of an outcome among a population are stable and where intervention effects are large and specific, such as was the case for the evaluation of the introduction of administrative restrictions on the use of certain antibiotics in Canada.[32] Evidence from before-after studies can be persuasive when assessing new behaviours, such as the use of new weaning food to promote child growth. However, improvements in growth or reductions in disease can only be attributed to the programme with confidence if data suggest other influences are unlikely. Problems may also arise from the incomparability of data collected before and after an intervention; for example, where the former relies on routine data while the latter involves evaluation surveys.

Table 1 Alternative design features

Optimal design feature	Alternative options ranked by internal validity of evidence	
	Options 1	**Options 2**
Pre-intervention random allocation of participants or clusters to intervention or control group	► Pre-intervention matching of control group to intervention group on key confounders, or statistical adjustment for baseline differences in these factors	► Restrict participation by key confounders ► Allocation by quantitative measure of need
Concurrent intervention and control group	► Comparison of changes over time in the intervention group with contemporaneous changes in the general population	► Before-and-after study with examination of other explanatory factors to increase the plausibility that any changes that are observed over time may be due to the intervention
Prospective follow-up from baseline to outcomes	► Proxy outcome measures ► Retrospective measurement of exposure	-

Estimating outcome rates in repeat cross-sectional surveys before and after an intervention (interrupted time-series study) may allow consideration of whether secular trends underlie observed changes. However, this may be expensive and also hampered by selection bias introduced by unmeasured changes in the composition of the sample over time. Such studies also cannot account for non-linear trends unless many pre-test measurements are taken, and may be insensitive to gradual changes such as might arise from anti-tobacco campaigns.[22]

None of the above strategies will address potential confounding from contemporaneous influential events (eg, a TV show with an HIV storyline). To strengthen their evidence, the study of pneumococcal conjugate vaccine analysed observed admission rates for a control condition (dehydration), which were very similar to the rates expected based on pre-intervention trends. This provided circumstantial evidence that the decline in pneumonia was not due to confounding from changes in healthcare coverage.[19] Alternatively, time-changes in the same outcome in a location or age group not subject to the intervention but subject to other factors that may cause changes over time could be examined. The pneumococcal conjugate vaccine study also considered changes in age-groups other than the infants who were the target of the intervention, showing reduced all-cause pneumonia admissions among adults aged 18–39 y (tentatively interpreted as evidence of a vaccine herd-effect among this group since they would include the parents of children directly exposed), but no evidence of declines among older adults. As with control of confounding, such strategies require knowledge of what other factors might influence outcomes. Process evaluations may also be useful in examining the plausibility of such influences.

A further problem, 'regression to the mean', can occur when participants are selected at baseline for their increased risk (eg, an HIV counselling intervention targets individuals concerned about their own risk), which then returns to a less-extreme level regardless of intervention.[33] Whether regression to the mean underlies apparent outcomes can be assessed by using a baseline measure that is the mean of several pre-intervention measures or, if this is unavailable, assessing whether intervention effects are apparent among participants at different levels of baseline risk.

ALTERNATIVES TO PROSPECTIVE FOLLOW-UP

It may be possible to use intermediate outcomes which predict final outcomes, but such proxies may underestimate or, more often, overestimate longer-term effects (because of the greater likelihood of dilution[34] rather than maintenance or multiplication[35] of effects). Another alternative is to use retrospective measures of exposure—for example, in case-control studies of vaccines.[36] Evaluations such as that of the Nepalese radio campaign employ a cross-sectional survey to compare outcomes between those reporting/not reporting exposure to the intervention. Because such studies rely on retrospective information they generally provide weak evidence about the temporal sequence of intervention and apparent outcomes. A further problem is that women reporting exposure to the intervention are likely to be a sub-set of those actually exposed and may differ in their response. Such studies are also vulnerable to confounding by unmeasured differences between those exposed/not exposed, a matter discussed further in our companion paper.

OTHER STRATEGIES

An additional way to assess whether outcomes arise from an intervention or other influences is to use process evaluation to determine whether there are plausible pathways linking intervention and outcome(s). While often used within RCTs,[37] these are even more important as a means of triangulation within non-randomised studies. For example, a non-randomised study of a teenage pregnancy prevention intervention drew on quantitative and qualitative data on young people's participation and negotiation skills to explore the plausibility of pathways to sexual health outcomes, as well as undertaking sensitivity analyses determining whether effects varied by exposure.[38] Examining the plausibility of causal pathways is facilitated by interventions being explicitly theorised.[39]

DISCUSSION

Based on discussion at a symposium involving individuals with experience of evaluating public health interventions, this paper has aimed to go beyond debates about the levels of evidence provided by RCT versus non-RCT studies in order to describe and exemplify practical solutions to problems encountered when choosing between designs. In summary, where genuine equipoise regarding health and other important outcomes exists, practical barriers are often surmountable so that RCTs (including stepped-wedge and cross-over trials) are possible. Where equipoise does not exist, or where delivery has already occurred or of necessity must be nationally consistent, RCTs and other prospective comparison designs may be unethical and/or unfeasible. In such cases, decisions should draw on other evidence as suggested for example in GRADE guidance.[5]

As we have seen, non-randomised studies can adopt strategies to reduce the possibility that other factors explain apparent intervention effects. Concurrent control groups are useful in minimising time-related confounding that hampers before-and-after studies and will be more convincing when evaluators take a comprehensive approach to identifying potential confounders.

What is already known on this subject

- Randomised trials are widely regarded as the 'gold standard' for estimating the causal effects of public health interventions.
- It has been argued that for certain decisions it is reasonable to infer intervention effects from non-randomised studies, and that dismissing designs other than the randomised trial might lead to a marginalisation of certain intervention types.
- However, existing reviews have not aimed to assess comprehensively the barriers and alternatives to randomised trials of public health interventions.

Process evaluations can check causal pathways and/or examine whether other factors might be influential. Checklists are useful for assessing the quality of non-randomised studies,[7 8] but case-by-case assessment of context-specific threats to validity and how to address these is also required.

Whether non-randomised studies actually provide useful evidence to guide decisions also depends on the effect sizes estimated and the decision/intervention being considered. Evidence of an intervention's impact will be more convincing when reported effect sizes are large since it is less likely that confounding or other sources of error can completely explain large effects[4 6] (although large effects arising from bias are not without precedent[40]). Evidence from non-randomised studies will also be more persuasive when results are consistent across studies[5 9] (although such studies can sometimes consistently mislead, exemplified by non-random studies of vitamin-A deficiency and mother-to-child HIV transmission mentioned above[30]).

When an intervention is costly, difficult to deliver or unacceptable to some stakeholders, when existing research suggests benefits may be small, or when there is evidence or scope for harmful effects, there is a strong argument that only an RCT will provide adequate evidence and barriers to undertaking one must be surmounted. However, where there exists evidence that an intervention is cheap and relatively easy to deliver, acceptable and there is minimal potential for harm, there is a stronger case for accepting evidence from other designs.[13 41] One relevant scenario here is confirmatory studies of the outcomes of intervention translated to new settings with few changes where previous RCTs report benefits and wider evidence suggests little scope for harm.[5]

Finally, further research is required on the effect of specific methods outlined above on the size and direction of bias in different areas of public health, as well as the use of STROBE and other guidelines to improve the reporting of non-randomised evaluations.

What this paper adds

- An RCT will generally not be possible where there is not equipoise regarding an intervention's health or other important effects, or where it must for practical, legal or bureaucratic reasons be delivered consistently across a nation.
- Evidence from non-randomised designs is more convincing when confounders are well-understood, measured and controlled, there is evidence for causal pathways linking intervention and outcomes and/or against other pathways explaining outcomes, and effect sizes are large.

Acknowledgements We would like to thank Diana Elbourne and Ben Armstrong for their contributions to the development of this paper. We would also like to thank those who attended a symposium on evaluating public health interventions convened by the London School of Hygiene and Tropical Medicine on 6 November 2006 for contributing insights and thus informing the development of this paper.

Funding The work was unfunded. JH is supported by a MRC/ESRC interdisciplinary postdoctoral fellowship.

Competing interests None.

Ethics approval Not required.

Contributors CB drafted the paper and reviewed the material that informs it. JH and SC also reviewed material informing the paper and contributed to drafting. DR, RH, MP and BK suggested examples and arguments for the paper and commented on successive drafts. All authors participated in writing this paper and have seen and approved the final version. CB had final responsibility for the decision to submit the paper for publication.

Provenance and peer review Not commissioned; externally peer reviewed.

REFERENCES

1. **Shadish WR,** Cook TD, Campbell DT. *Experimental and quasi-experimental designs for generalized causal inference*. Boston, MA: Houghton Mifflin, 2002.
2. **Campbell DT,** Russo JJ. *Social experimentation*. Thousand Oaks, CA: Sage, 1999.
3. **Habicht JP,** Victora CGV, Vaughan JP. Evaluation designs for adequacy, plausibility and probability of public health programme performance and impact. *Int J Epidemiol* 1999;**28**:10—18.
4. **Victora C,** Habicht JP, Bryce J. Evidence-based public health: moving beyond randomized trials. *Am J Public Health* 2004;**94**:400—05.
5. **GRADE Working Group.** Grading quality of evidence and strength of recommendations. *BMJ* 2004;**328**:1490.
6. **West SG,** Duan N, Pequegnat W, *et al*. Alternatives to the randomized controlled trial. *Am J Public Health* 2008;**98**:1359—66.
7. **von Elim E,** Altman DG, Egger M, *et al*. Strengthening the reporting of observational studies in epidemiology (STROBE) statement: guidelines for reporting observational studies. *BMJ* 2007;**335**:806—08.
8. **Des Jarlais DC,** Lykes C, Crepaz C, *et al*. Improving the reporting quality of non-randomized evaluations of behavioural and public health interventions: the TREND statement. *Am J Public Health* 2004;**94**:361—65.
9. **Kirkwood BR,** Cousens SN, Victora CG, *et al*. Issues in the design and interpretation of studies to evaluate the impact of community-based interventions. *Tropical Med Int Health* 1997;**2**:1022—29.
10. **Brewin CR,** Bradley C. Patient preferences and randomised controlled trials. *BMJ* 1989;**299**:313—15.
11. **Bonell C,** Hargreaves JR, Strange V, *et al*. Should structural interventions be evaluated using RCTs? The case of HIV prevention. *Soc Sci Med* 2007;**63**:1135—42.
12. **Cuijpers P,** Jonkers R, de Weerdt I, *et al*. The effects of drug abuse prevention at school: the 'Healthy School and Drugs' project. *Addiction* 2002;**97**:67—73.
13. **Thomson H,** Hoskins R, Petticrew M, *et al*. Evaluating the health effects of social interventions. *BMJ* 2004;**328**:282—85.
14. **Torgerson D,** Sibbald B. Understanding controlled trials: What is a patient preference trial? *Br Med J* 1998;**316**:360.
15. **Toroyan T,** Roberts I, Oakley A, *et al*. Effectiveness of out-of-home day care for disadvantaged families: randomised controlled trial. *BMJ* 2003;**327**:906—10.
16. **Armstrong Schellenberg JR,** Adam T, Mshinda H, *et al*. Effectiveness and cost of facility-based Integrated Management of Childhood Illness (IMCI) in Tanzania. *Lancet* 2004;**364**:1583—94.
17. **Bonell C,** Oakley A, Hargreaves J, *et al*. Trials of health interventions and empirical assessment of generalizability: suggested framework and systematic review. *BMJ* 2006;**333**:346—49.
18. **Hussey MA,** Hughes JP. Design and analysis of stepped wedge cluster randomized trials. *Contemp Clin Trials* 2007;**28**:182—91.
19. **Grijalva CG,** Nuorti JP, Arbogast PJ, *et al*. Decline in pneumonia admissions after routine childhood immunisation with pneumococcal conjugate vaccine in the USA: a time-series analysis. *Lancet* 2007;**369**:1179—86.
20. **Hutchinson P,** Wheeler J. Advanced methods for evaluating the impact of family planning communication programs: evidence from Tanzania and Nepal. *Stud Fam Plann* 2006;**37**:169—86.
21. **Black N.** Why we need observational studies to evaluate the effectiveness of health care. *BMJ* 1996;**312**:1215—18.
22. **Susser M.** Some principles in study design for preventing HIV transmission: rigour or reality. *Am J Public Health* 1996;**86**:1713—16.
23. **Cowan F,** Plummer M. Biological, behavioural and psychological outcome measures. In: Stephenson J**,** Imrie J, Bonell C, eds. *Effective sexual health interventions: issues in experimental evaluation*. Oxford: Oxford University Press, 2003:111—36.
24. **Snapinn SM.** Noninferiority trials. *Curr Control Trials Cardiovasc Med* 2000;**1**:19—21.
25. **McNeely C,** Falci C. School connectedness and the transition into and out of health-risk behavior among adolescents: a comparison of social belonging and teacher support. *J Sch Health* 2004;**74**:284—92.

26. **Campbell DT,** Boruch RF. Making the case for randomized assignments to treatments by considering the alternatives: six ways in which quasi-experimental evaluations in compensatory education tend to underestimate effects. In: Bennett CA**,** Lumsdaine A, eds. *Evaluation and experimentation: some critical issues in assessing social programs*. New York, NY: Academic Press, 1975: 195—296.
27. **Deeks JJ,** Dinnes J, D'Amico R, *et al*. Evaluating non-randomised intervention studies. *Health Technol Assess* 2003;**7**:1—187.
28. **Petticrew M.** Why certain systematic reviews reach uncertain conclusions. *BMJ* 2003;**32**:756—58.
29. **Ioannidis JPA,** Haidich A, Pappa M, *et al*. Comparison of evidence of treatment effects in randomized and non-randomized studies. *J Am Med Assoc* 2001;**286**:821—30.
30. **Dreyfuss ML,** Fawzi WW. Micronutrients and vertical transmission of HIV-1. *Am J Clin Nutr* 2002;**75**:959—70.
31. **Egan M,** Petticrew M, Ogilvie D, *et al*. New roads and human health: a systematic review. *Am J Public Health* 2003;**93**:1463—71.
32. **Marshall D,** Gough J, Grootendorst P, *et al*. Impact of administrative restrictions on antibiotic use and expenditure in Ontario: time series analysis. *J Health Serv Res Policy* 2006;**11**:13—20.
33. **Sacks H,** Chalmers TC, Smith HJ. Randomized versus historical controls for clinical trials. *Am J Med* 1982;**72**:233—40.
34. **Faggiano F,** Vigna-Taglianti FD, Versino E, *et al*. School-based prevention for illicit drugs use. *Cochrane Database Syst Rev* 2005, 2;CD003020. DOI@: 10.1002/14651858.
35. **Weikart DP,** Berrueta-Clement JR, Schweinhart LJ, *et al*. *Changed lives: the effects of the perry pre-school program on youths through age nineteen*. Ypsilanti, MI: High/Scope Press, 1984.
36. **Rodrigues LC,** Smith PG. Use of the case-control approach in vaccine evaluation: efficacy and adverse effects. *Epidemiol Rev* 1999;**21**:56—72.
37. **Oakley A,** Strange V, Bonell C, *et al*. Integrating process evaluation in the design of randomised controlled trials of complex interventions: the example of the RIPPLE Study. *BMJ* 2006;**332**:413—16.
38. **Wiggins M,** Bonell C, Burchett H, *et al*. *Young people's developmentpProgramme final report*. London: Institute of Education, 2008.
39. **Rychetnik L,** Frommer M, Hawe P, *et al*. Criteria for evaluating evidence on public health interventions. *J Epidemiol Community Health* 2002;**56**:119—27.
40. **Vessey MP,** Yeates D, Flavel R, *et al*. Pelvic inflammatory disease and the intrauterine device: findings in a large cohort study. *BMJ (Clin Res Ed)* 1981;**282**:855—7.
41. **Ross DA,** Wight D, Dowsett G, *et al*. *The weight of evidence: a methodology for assessing the strength of evidence on the effectiveness of HIV prevention interventions among young people*. Geneva: WHO, 2006.