Alternatives to randomisation in the evaluation of public-health interventions: statistical analysis and causal inference

S Cousens, J Hargreaves, C Bonell, B Armstrong, J Thomas, B R Kirkwood, R Hayes

Alternatives to randomisation in the evaluation of public-health interventions: statistical analysis and causal inference

Cousens S[1], Hargreaves JR[1], Bonell C[1], Armstrong B[1], Thomas J[2], Kirkwood BR[1], Hayes R[1].

[1]London School of Hygiene and Tropical Medicine

[2] EPPI Centre, Social Science Research Unit, Institute of Education, University of London

Address for correspondence:

Simon Cousens

Infectious Disease Epidemiology Unit

London School of Hygiene and Tropical Medicine

Keppel Street

London WC1E 7HT

Tel: +44 (0) 20 7927 2422

Fax: +44 (0) 20 7636 8739

Email: simon.cousens@lshtm.ac.uk

Words -	3077
Summary -	246
References -	37
Tables -	1
Figures -	4

Abstract

Background

In non-randomised evaluations of public-health interventions statistical methods to control confounding will usually be required. We summarise a variety of approaches to the control of confounding, highlighting key issues and the assumptions on which these methods are based.

Method

Explanatory review.

Results

To control confounding, standard stratification and regression techniques will often be appropriate but propensity scores may be useful where many confounders need to be controlled and data are limited. All these techniques require that key putative confounders are measured accurately. Instrumental variables offer, in theory, a solution to the problem of unknown or unmeasured confounders but identifying an instrument which meets the required conditions will often be challenging. Obtaining measurements of the outcome variable in both intervention and control groups before the intervention is introduced allows balance to be assessed, and these data may be used to help control confounding. However, imbalance in outcome measures at baseline poses challenges for the analysis and interpretation of the evaluation, highlighting the value of adopting a design strategy that maximizes the likelihood of achieving balance. Finally, when it is not possible to have any concurrent control group, making multiple measures of outcome

pre- and post-intervention can enable the estimation of intervention effects with appropriate statistical models.

Conclusion

For non-randomised designs careful statistical analysis can help reduce bias by confounding in estimating intervention effects. However, investigators must report their methods thoroughly and be conscious and critical of the assumptions they must make whenever they adopt these designs.

Alternatives to randomisation in the evaluation of public-health interventions: statistical analysis and causal inference

Introduction

In our previous paper, we discussed barriers to randomised controlled trials (RCTs) of public-health interventions and suggested alternative design strategies.[1] In this paper, we discuss the statistical analysis of data from non-randomised evaluations, dealing particularly with *confounding*. We outline key issues and options available when planning analyses of these data. In practice, the most appropriate statistical approach will differ from case to case, but transparent description of the design and analysis process is essential [2, 3], as for RCTs [4].

Causal effects and confounding in non-randomised evaluations of public-health interventions

Evaluations of public-health interventions aim to estimate the *causal effect* of an intervention by which we mean a quantitative measure of the difference between the level of the outcome had everybody in the population of interest been exposed and the level of the outcome had everybody been unexposed. Readers are referred to Hernan (2004) for a fuller discussion of causal effects in epidemiology.[5] Our emphasis on public-health interventions also implies we are most concerned with estimating overall (population-level) effects of intervention strategies, combining direct and indirect effects.[6]

RCTs aim to achieve *balance* between treatment and non-treatment groups, meaning that these groups are alike with regard to all factors that might influence the outcome measure (both known and unknown), other than exposure to the intervention of interest. RCTs achieve balance by using chance to determine which people or units receive an intervention, and by applying this process many times over (i.e. allocating many units to the two groups).[7]

Most non-RCT designs also seek to achieve balance.[1] Nevertheless, *imbalance* is more likely with these designs and consequently they risk generating a *confounded* effect estimate: one that mixes the effect of the intervention with other causal effects.[8]

Consider a study to evaluate the effect of a radio soap-opera designed to encourage contraceptive use in Nepal using data collected from over 8,000 women in a cross-sectional survey (see box 1 in our previous paper).[1, 9] The authors compared contraceptive use among women who reported listening to the programme in the previous six months with that among women who did not. An *unadjusted* analysis found that the prevalence of contraceptive use was 12% higher among listeners than non-listeners (Table 1). However, it would be premature to conclude that this reflects an intervention effect. While this study suffered from many potential biases (as do most evaluations) we highlight here the issue of *confounding*. For example, level of education might differ between those who did and did not listen to the soap-opera, and might also, independently, influence contraceptive use (Figure 1). Of course, educational level is

only one of many potential confounders, and a more complex causal diagram than Figure 1 could be drawn to help identify other variables that should be controlled.[10]

Statistical methods for controlling confounding

(i) Stratification and regression

In the traditional approach to controlling confounding, women would be grouped ("stratified") according to their educational level, for example "none", "attended primary school" and "completed primary school". The association between intervention and outcome is examined within each group. Any association within groups cannot be due to confounding by educational status because women in each group have the same level of education, assuming this is correctly measured. If the intervention effect is approximately the same in all sub-groups, a weighted average of the stratum-specific estimates provides an adjusted effect estimate free of confounding by that variable.

Regression modelling can include multiple confounding factors as explanatory variables[11], and was used in the Nepal study to control 11 potential confounders, assuming no effect modification. This *adjustment* reduced the estimated effect from +12% to +6% (Table 1), consistent with the *unadjusted* estimate being partly confounded (overestimating the true effect). However, this adjusted estimate is unconfounded only if all important confounders are identified and measured accurately[12, 13]. Since we can rarely, if ever, verify this, residual confounding remains a concern.

(ii) *Propensity scores*

Propensity scores are another more recent approach to controlling confounding.[14] The following steps are taken:

1. A regression model is used to identify factors that 'predict' exposure to the intervention. The model is used to calculate each individual's predicted probability of, rather than actual, exposure to the intervention (e.g. "listened to the soap opera").

2. Individuals with similar propensity scores are grouped. Within each group, some individuals will actually have been exposed to the intervention and some not. Since individuals in each group had the same propensity to be exposed, the method assumes that actual exposure within these groups was random.

3. Stratified analysis can compare outcomes between exposed and unexposed individuals within each propensity-score group and by including the propensity score group in regression analysis it is possible to obtain an unconfounded estimate of the intervention effect. Alternatively, each exposed individual may be matched with an unexposed individual with the same or similar propensity score and the analysis restricted to these pairs.

When used in the Nepal study, this approach yielded an effect-size estimate of a 9% increase in contraceptive use. Unfortunately, no details were provided of the variables used to estimate the propensity scores though this choice may affect the estimate obtained. Future evaluations adopting this approach should provide these details [2, 3].

Propensity scores are increasingly popular.[15] One advantage is that their use can reduce the number of parameters (i.e. variables and categories within these) to be estimated in a regression model. When the number of parameters is large relative to the data available, estimates can become biased and confidence intervals unreliable. In the current example, using propensity scores rather than standard regression reduced the number of parameters from 25 to 8. In this case a model with 25 parameters should not have introduced bias,[16] since over 2500 women reported using contraceptives. However, in smaller evaluations or those focused on rare outcomes, propensity-score approaches can reduce such problems, [14, 17] and may be more robust than standard regression when the number of events-per-confounder is small (< 8).[18]

Problems remain however. In practice, comparisons between regression and propensity-score methods suggest they usually yield similar results.[15, 19, 20] Like standard regression methods, a propensity-score analysis faces the problem of unmeasured or poorly-measured variables since all important predictors of exposure that are causally associated with the outcome must be included. Furthermore, using statistical methods to identify 'predictors' of exposure without considering underlying causal relationships

might be problematic. Investigators should not include 'predictors' which are in fact consequences of exposure, since this will lead to 'over-adjusted' models and biased effect estimates.[10, 21] Alternatively, if multiple predictors of exposure that are not causally associated with the outcome are included, then power may be unnecessarily sacrificed since, as with other methods of controlling confounding, including multiple factors in a model can reduce power.[22]

(iii) Instrumental variables

Instrumental variables are also increasingly popular, purportedly removing the need to identify and measure all potential confounders.[23, 24] This approach requires an "instrument" that meets the following conditions (Figure 2):

1. It is a cause or proxy for a cause of exposure to the intervention;

2. It is not a cause of the outcome other than through the intervention; and

3. It is not associated with any unmeasured confounders of concern in the study population.

Identifying an instrument that satisfies these conditions allows us to generate an unconfounded estimate of the intervention effect ('effect A' in Figure 2), by comparing the magnitude of the association between the instrument and the outcome ('effect C')

with that between the instrument and exposure to the intervention ('effect B'). In the case of the Nepal example, the intervention effect is estimated by dividing the effect (risk difference) of the instrument on the outcome by the effect (risk difference) of the instrument on the exposure. The precise manner in which estimates are calculated differs depending on the situation.[23, 25]

Instrumental-variable analysis may provide different estimates of treatment effect to standard or propensity-score methods.[20] In the Nepal example, "listens to radio weekly" was used as an instrument (see figure 2). This analysis suggested an 8.5% increase in contraceptive use associated with the intervention. However, we must question whether conditions 2 and 3 were met. First, listening to the radio weekly (including other programmes) could conceivably have a direct effect on contraceptive practices (violating 2). Second, important unmeasured confounders (e.g. social and cultural factors) might also be associated with this instrument (violating 3). This illustrates that it is often difficult to identify an appropriate instrument. Furthermore, we can never empirically verify the conditions required for a valid instrument. Consequently, "the fundamental problem of causal inference from observational data – the reliance on assumptions that cannot be empirically verified — is not solved but simply shifted to another realm".[23] Finally, even if all three conditions above are met, if the correlation between the instrument and the intervention is not strong (the instrument is "weak"), the standard error of the intervention effect will be large and the confidence interval for the effect will be wide.[26]

(iv) Using pre-intervention measures of the outcome variable in analysis

In both randomised and non-randomised evaluations, obtaining measures of the outcome variable prior to the introduction of the intervention can be useful to explore *balance.* These data can also be used to control confounding since they reflect the prior influence of multiple factors that influence the outcome in the absence of intervention. The data can be incorporated into analysis in two ways:

1. Treat them in the same way as other potential confounders and fit a regression model in which the pre-intervention measure of the outcome is included alongside other potential confounders.

2. Calculate the change in the outcome and base the analysis on the difference in the changes in the two groups.

If intervention and control groups are similar with respect to baseline measures of the outcome (i.e. there is *balance*), or where small differences arise by chance (as could also happen in an RCT), both approaches provide unbiased estimates of the intervention effect but regression will provide a more precise estimate and is therefore preferred.[27]

However, when non-random allocation results in two groups which are drawn from two different populations, and hence are unbalanced at baseline, the two approaches can give contradictory results. This has been described as “Lord’s paradox” and was first

identified in the context of individual-level data.[27-29] To illustrate this paradox with an example relevant to our purposes, consider a hypothetical cluster-allocated, non-randomised study of an intervention aiming to lower mean systolic blood pressure (MSBP) among individuals in workplaces by influencing exercise and smoking (100 intervention, 100 control sites). Intervention allocation was non-random, being determined by stakeholder meetings to identify which workplaces received the intervention.[1] This resulted in a systematic imbalance in MSBP between the two arms at baseline (intervention 120mmHg, control 115mmHg). Following intervention, measures of MSBP were taken from individuals in all workplaces. MSBP did not change between baseline and follow-up in either the control or intervention workplaces.

In this situation, because blood pressure differed between the groups (was *unbalanced*) to begin with, regression analysis can incorrectly suggest that there was an effect of the intervention in some situations, while an analysis of change in the outcome does not.[27-29] Here we offer one suggestion for how this paradox might occur and reflect on guidance for evaluators in this situation. Figure 3 shows the results of two simulations of the experiment with pre-intervention MSBP plotted against MSBP post-intervention for all intervention and control workplaces. In Figure 3a, we assume that MSBP was measured "perfectly". Over 1000 simulations there was no evidence for a difference between the groups from either regression analysis or analysis of changes. However, when we allowed "noise" in the baseline measurements of MSBP (Figure 3b), there was evidence of a difference between the groups in regression analysis but not change-scores**.** Our simulation assumed that each individual had the same true underlying systolic blood

pressure at both baseline and follow-up, i.e. we assume no real change in either group, but that at each time-point there was independent random measurement error. Our simple simulation suggests that one possible explanation for paradoxical findings when comparing an analysis of changes with a regression analysis is that “noise” results in dilution of the association between pre- and post-intervention measures as identified by linear regression. This is known as regression-dilution bias.[30] This phenomenon reduces the gradient of the regression lines in Figure 3b compared to Figure 3a. The regression lines for each group are consequently shallower, but are also centred on different means (because of the baseline imbalance) resulting in a vertical gap between the two lines. In regression analysis, this gap is equivalent to the estimated parameter for intervention effect, and might then be incorrectly interpreted in this way.

Further research is necessary to characterise the statistical properties associated with this phenomenon; we have offered only a simplified illustration. Such work is necessary because this situation might plausibly arise in non-randomised evaluations of public-health interventions. For example, we identified an ongoing evaluation of the impact of introducing youth-centres and ‘adolescent-friendly’ clinics on HIV prevalence among 15-24 year-olds in South-African communities. The centres were purposively placed in disadvantaged areas for strategic purposes.[31] The evaluation design aims to compare future HIV prevalence through surveys in 11 youth-centres, 11 clinics and 11 control sites but baseline HIV prevalence was higher in areas where youth-centres were placed (15.7%) than controls (13.8%: adjusted OR = 1.41 95% CI 0.96, 2.07).[31] It is inevitable that there is imprecision in these site-specific estimates since they are based on a sample

of the population. Given this imbalance at baseline, a future regression analysis using data from a new sample post-intervention is likely to be biased by the regression-dilution bias illustrated above. However, for change-scores to produce a valid estimate of the intervention effect, we must assume that the intervention has a constant additive effect regardless of the level at baseline on whatever scale the analysis is performed (for example, log (odds) in the case of logistic regression). The key message of Lord's paradox is, therefore, that when non-randomised intervention and control groups are unbalanced at baseline, any attempt at causal inference will be fraught with difficulty.[27, 29]

(v) Imbalance and inference when the number of units studies is small

Public-health interventions are often delivered to 'clusters' of people and for practical reasons the number of clusters included in an evaluation is sometimes small.[1] Cluster allocation must be appropriately taken into account in analyses, and this is relevant to both randomised and non-randomised designs.[32] Low statistical power in such studies, including those where large numbers of individuals but only a small number of clusters are enrolled, is a major barrier to statistical inference. We do not seek to review this issue here, other than to identify that allocation of a small number of units may also be a reason why imbalance might arise in non-randomised evaluations and thus indicate the need for control of confounding. However, many relevant statistical methods require additional assumptions when cluster numbers are low. We thus re-iterate the more general point that while in-depth studies of interventions delivered in a few study units can provide

valuable information on process and provide some forms of evidence to inform public-health decision-making [33], they are highly constrained in their capacity to provide quantitative estimates of intervention effect.

(vi) Time-series studies

When it is not possible to recruit a concurrent comparison group it may instead be possible to compare each unit pre- and post-intervention. However, once again, a "fair comparison" should be made. The before/after approach, and more sophisticated variants of this in which multiple measures of outcome are made over time, controls for sources of confounding that are static over time but not time-varying factors such as maturational, seasonal or secular trends.

"Interrupted time-series analysis" requires data on multiple measures of the outcome pre- and post-intervention.[34] The following steps are taken:

1) The extent of variation in the outcome over time due to factors other than the intervention (e.g. seasonal trends) [35] is estimated statistically.

2) A statistical model is used to predict the "expected" outcome at the end of the intervention period had the intervention not been delivered.

3) This “expected value” is compared with the “observed” post-intervention level to determine the intervention “effect”.

For example, the effect of introducing pneumococcal conjugate vaccine (PCV-7) for infants in the USA was evaluated by examining monthly pneumonia admissions (see previous paper, box 1).[1, 36] There were significant seasonal variation in trends in admission (Figure 4). An expected admission rate at the end of the intervention period was obtained by extrapolating available trend data after modelling seasonal fluctuations. The analysis found that the seasonally-corrected admission rate by December 2004 was 39% lower than the expected rate (95% CI 22%-52%) providing an estimate of the effect of PCV-7 (Figure 4a).

Interrupted–time-series analysis provides better estimates than simple before/after studies as long as putative time-varying confounding factors are measured and modelled.[37] Acute effects of rapid introduction of an intervention are generally easiest to differentiate from other sources of variation in time-series analyses.[34] Difficulties arise if an intervention is implemented gradually or has a long latent interval before exerting effect (e.g. the effect of anti-smoking campaigns on lung-cancer rates). A challenge also arises in deciding how complicated a trend to allow for in estimating expected values post-intervention. Non-linear trends can be modelled and provide better control of confounding, but require data for many time-points and may result in less-precise effect estimates. Furthermore, even after modelling trends and allowing for seasonality, there may be auto-correlation between outcome levels at adjacent time-points.[38] This auto-

correlation can lead to over-estimating the precision of intervention effects and provide confidence intervals that are too narrow. For continuous outcomes, there are techniques available to take account of this; for counts, a little ingenuity is required.[39]

Conclusion

Non-randomised evaluations are essential to inform public-health decision making where there are clear barriers to the conduct of RCTs. Over two papers, we have discussed design and analysis choices in order to ensure a "fair comparison" is made. Confounding, however, remains a major concern in these studies and investigators will face more complex problems even than those we have discussed here such as dealing with covariates that change over time.[40] Evaluators and analysts have various options to consider but must make careful, informed choices that fit their context. We hope to have aided these choices. Most importantly, as we and others have stressed, investigators should transparently outline the steps taken in design and analysis so that others can judge the value of the estimates produced.

Tables and Figures

Figure 1. Simple causal diagram of potential causal effects and links between exposure to the intervention, educational level and contraceptive use

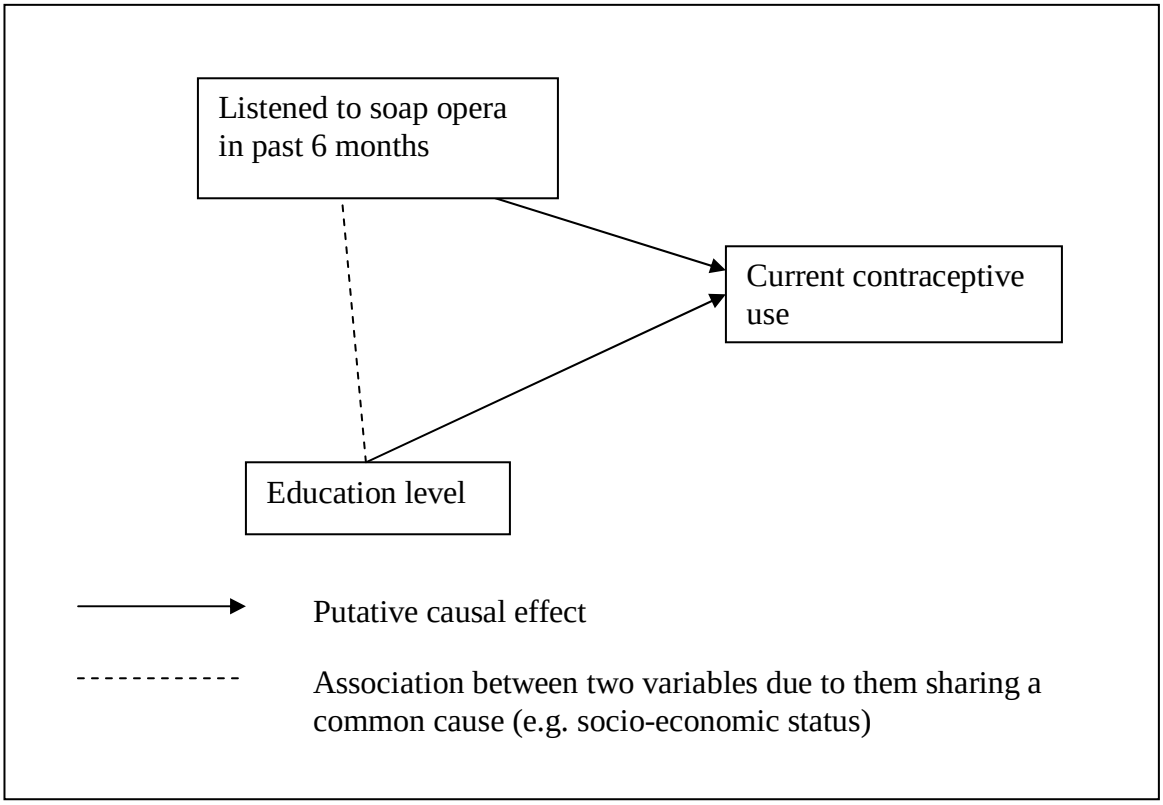


Table 1. Results of an evaluation of a family-planning communication intervention in Nepal[+]

	Group		Estimated effect of the intervention			
	Not listened to soap opera	Listened to soap opera	Unadjusted analysis	Including potential confounders in regression model[*]	Using propensity scores approach[^]	Using "listens to the radio weekly" as an instrument
Current use of a modern contraceptive method	31%	43%	+12%	+6.2%	+9.2%	+8.5%

Notes on table: [+] The data on which these analyses are based are downloadable on request from http://www.measuredhs.com/.

* Probit regression, a statistical approach from the same family as logistic and linear regression that are more commonly used in public-health, was used to estimate risk differences as presented in the table. The principles we discuss apply equally to other regression methods. Variables included in the model as confounders were: woman's age, woman's educational level, household asset index, religion, lives in rural area, visited by family-planning worker in past 12 months, currently employed, number of living children, husband's education, watches television weekly, listens to radio weekly.

^ No information provided on variables used to calculate propensity score

Figure 2. Causal diagram indicating the conditions under which a variable (the "instrument") will enable the investigator to control unmeasured confounding variables

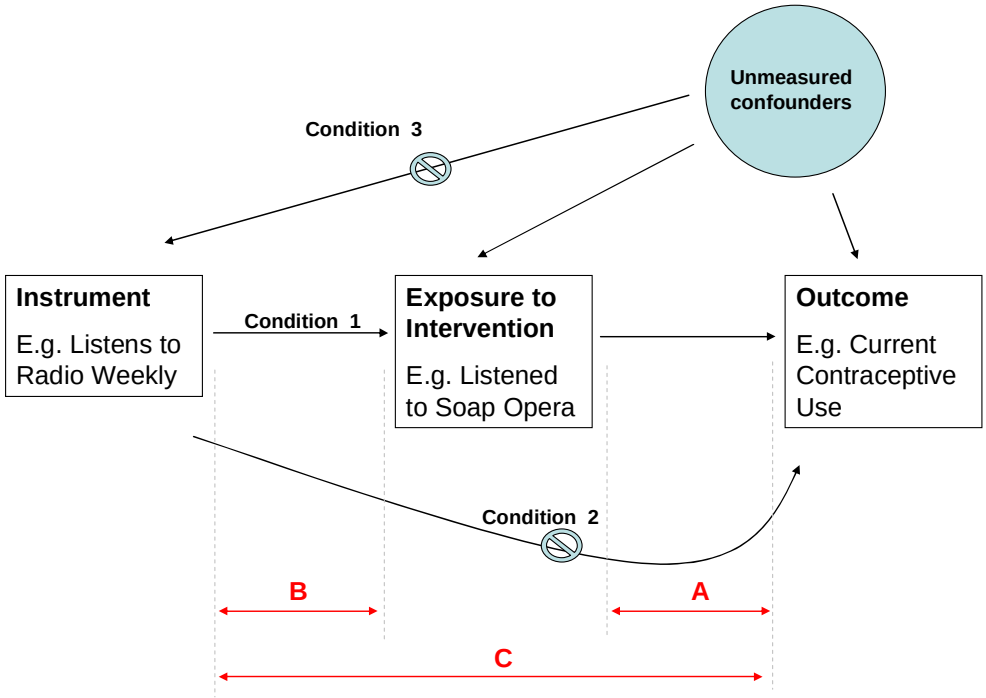
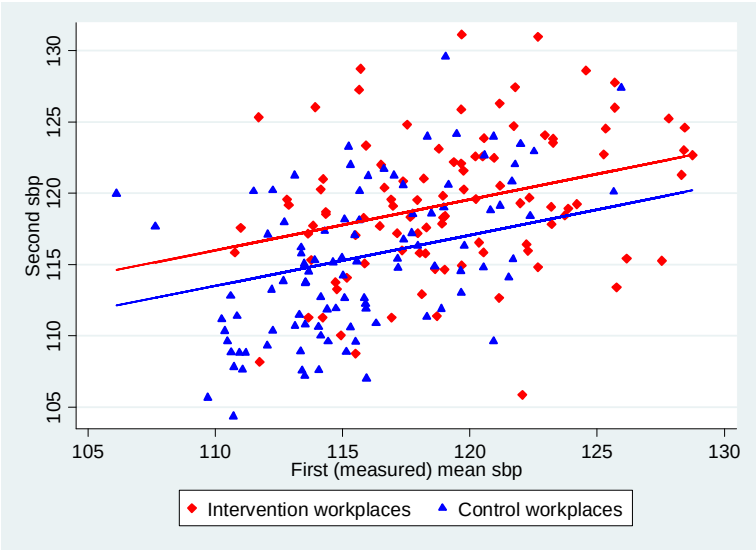


Figure 3. Simulated data illustrating possible mechanism for "Lord's paradox". The vertical distance between the two regression lines corresponds to the "intervention effect" estimate obtained from regression. In figure 3a, with no "noise" in the baseline measure, this suggests no intervention effect, consistent with the analysis of change-scores. In figure 3b, with "noise" in the measurements, regression suggests an intervention effect.

(a)

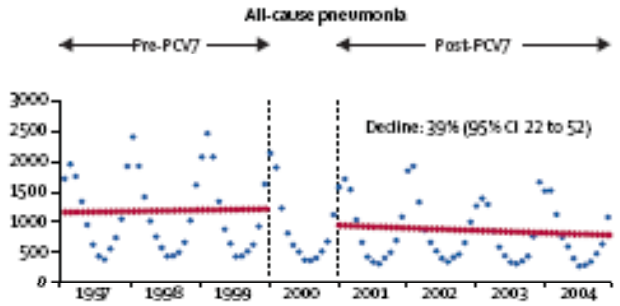
(b)



Details of simulation: Using STATA version 10, a simulated population of 100 intervention clusters were drawn from a sampling distribution with mean of 120, and standard deviation of 3 and 100 control clusters were drawn from a population with mean 115, standard deviation 3 to represent baseline values. A standard deviation of 3mmHg in cluster level mean SBPs could plausibly arise in an evaluation – for example if the standard deviation of individual blood pressure values in each population was 18mmHg and the number of individuals enrolled in each cluster was 36. Follow up measures simulating random variation over time but no intervention effect (Figure 3a) were simulated by adding an additional random value (mean 0, standard deviation 3) to each baseline value. "Noise" at baseline and follow up (in Figure 3b) was simulated by adding a further random value to both baseline and follow-up measurements (mean 0 , standard deviation 3). For both Figure 3a and 3b, the individual red and blue points represent the cluster-level mean values from a single simulation, while the lines from regression represent the results of 1000 repeated simulations of the experiment. In Figure 3a with no noise the black line indicates that the two regression lines are coincident; in Figure 3b the red and blue lines are separated by a distance of 2.5, representing the estimated average "intervention effect" in regression analysis, and illustrating one possible mechanism for Lord's paradox.

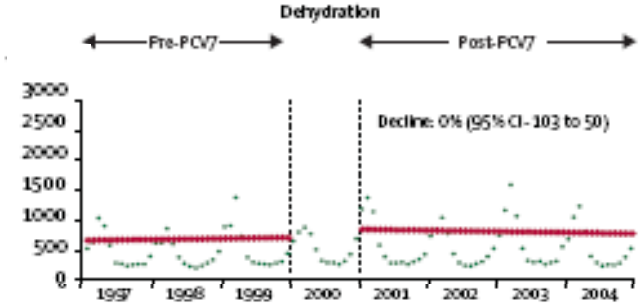
Figure 4. Trends in monthly US admission rates (1997–2004) for (a) all-cause pneumonia and (b) dehydration (control condition) among under 2 year olds before and after routine immunisation of children with PCV7 (partially reproduced from Grijalva et al.[36])

a.



Estimated rate of admission/ 100,000	Expected rate of admission / 100,000	Rate difference (95% CI)	Proportional reduction in admission rate (95% CI)
790.9	1296.9	-505.9 (-291.4,-674.7)	39% (22%,52%)

b.



Estimated rate of admission/ 100,000	Expected rate of admission / 100,000	Rate difference (95% CI)	Proportional reduction in admission rate (95% CI)
778.9	775.3	3.6 (797.2, –389.5)	-0.5% (-,50%)

What we already know on this subject

Non-randomised evaluations of public health interventions are more likely than randomised designs to suffer from imbalance between intervention and control groups. Consequently they risk generating confounded effect estimates that mix the effect of the intervention with other causal effects. A variety of statistical approaches are available to minimise confounding but existing reviews have not discussed these in a way that can help guide the planning of evaluations of public health interventions where randomisation is not possible.

What this paper adds

Like regression and stratification techniques, propensity scores require that key putative confounders are measured accurately, but may be particularly useful where many confounders need to be controlled and data are limited. Instrumental variables offer, in theory, a solution to the problem of unknown or unmeasured confounders but identifying a valid instrument will often be challenging. Outcome measures taken at baseline can be used in analysis but imbalance poses complex challenges for analysis and interpretation. Finally, time-series analysis can enable the estimation of intervention effects with appropriate, but complex, statistical models.

Acknowledgments

We would like to thank those who attended a symposium on evaluating public-health interventions convened by the London School of Hygiene and Tropical Medicine on 6th November 2006 for contributing insights and thus informing the development of this paper.

Authors' contributions

S Cousens drafted the paper and reviewed the material which informs it. J Hargreaves, C Bonell, R Hayes, J Thomas and B Armstrong also reviewed material informing the paper and contributed to the drafting of it. B Kirkwood suggested examples and arguments for the paper and commented on a draft. All authors participated in writing this paper and have seen and approved the final version. S Cousens had final responsibility for the decision to submit the paper for publication.

Conflict of interest and funding

No authors declare any conflicts of interest regarding this paper. The work was undertaken unfunded. JH is supported by an MRC/ESRC interdisciplinary postdoctoral fellowship.

Ethics

Since the article is based on a review of existing literature, there was no requirement to seek ethical review or human subjects' informed consent to participate.

References

1 Bonell CP, Hargreaves JR, Cousens SN, *et al.* Evaluating the effects of public-health interventions: barriers to randomized trials and alternative options. *BMJ* 2008;**Under Review**.

2 Des Jarlais DC, Lyles C, Crepaz N. Improving the reporting quality of nonrandomized evaluations of behavioral and public health interventions: the TREND statement. *Am J Public Health* 2004;**94**:361-6.

3 von Elm E, Altman DG, Egger M, *et al.* The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. *PLoS Med* 2007;**4**:e296.

4 Altman DG. Better reporting of randomised controlled trials: the CONSORT statement. *Bmj* 1996;**313**:570-1.

5 Hernan MA. A definition of causal effect for epidemiological research. *J Epidemiol Community Health* 2004;**58**:265-71.

6 Halloran ME, Haber M, Longini IM, Jr., *et al.* Direct and indirect effects in vaccine efficacy and effectiveness. *Am J Epidemiol* 1991;**133**:323-31.

7 Bradford Hill A. The Environment and Disease: Association or Causation? *Proc R Soc Med* 1965;**58**:295-300.

8 Rothman KJ, Greenland S, Lash TL. *Modern Epidemiology*. New York: Lippincott, Williams & Wilkins 2008.

9 Hutchinson P, Wheeler J. Advanced methods for evaluating the impact of family planning communication programs: evidence from Tanzania and Nepal. *Studies in Family Planning* 2006;**37**:169-86.

10 Greenland S, Pearl J, Robins JM. Causal diagrams for epidemiologic research. *Epidemiology* 1999;**10**:37-48.

11 Kirkwood B, Sterne J. *Essential Medical Statistics, 2nd Edition*. Oxford: Blackwell Publishing 2003.

12 Fewell Z, Davey Smith G, Sterne JAC. The Impact of Residual and Unmeasured Confounding in Epidemiologic Studies: A Simulation Study. *American Journal of Epidemiology* 2007;**166**:646-55.

13 Kupper LL. Effects of the use of unreliable surrogate variables on the validity of epidemiologic research studies. *Am J Epidemiol* 1984;**120**:643-8.

14 D'Agostino RB, Jr. Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. *Stat Med* 1998;**17**:2265-81.

15 Stürmer T, Joshi M, Glynn RJ, *et al.* A review of the application of propensity score methods yielded increasing use, advantages in specific settings, but not substantially different estimates compared with conventional multivariable methods. *Journal of Clinical Epidemiology* 2006;**59**: 437-47.

16 Peduzzi P, Concato J, Kemper E, *et al.* A simulation study of the number of events per variable in logistic regression analysis. *Journal of Clinical Epidemiology* 1996;**49** 1373-9.

17 Rosenbaum PR, Rubin DB. Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *American Statistician* 1985;**39**:33-8.
18 Cepeda MS, Boston R, Farrar JT, *et al.* Comparison of Logistic Regression versus Propensity Score When the Number of Events Is Low and There Are Multiple Confounders *American Journal of Epidemiology* 2003;**158**:280-7.
19 Shah BR, Laupacis A, Hux JE, *et al.* Propensity score methods gave similar results to traditional regression modeling in observational studies: a systematic review. *Journal of Clinical Epidemiology* 2005;**58**:550-9.
20 Stukel TA, Fisher ES, Wennberg DE, *et al.* Analysis of observational studies in the presence of treatment selection bias: effects of invasive cardiac management on AMI survival using propensity score and instrumental variable methods. *Jama* 2007;**297**:278-85.
21 Hernan MA, Hernandez-Diaz S, Werler MM, *et al.* Causal knowledge as a prerequisite for confounding evaluation: an application to birth defects epidemiology. *Am J Epidemiol* 2002;**155**:176-84.
22 Winkelmayer WC, Kurth T. Propensity scores: help or hype? *Nephrol Dial Transplant* 2004;**19**:1671-3.
23 Hernan MA, Robins JM. Instruments for causal inference: an epidemiologist's dream? *Epidemiology* 2006;**17**:360-72.
24 Angrist JD, Imbens GW, Rubin DB. Identification of causal effects using instrumental variables. *Journal of the American Statistical Association* 1996;**91**:444-55.
25 Martens EP, Pestman WR, de Boer A, *et al.* Instrumental variables: application and limitations. *Epidemiology* 2006;**17**:260-7.
26 Newhouse JP, McClellan M. Econometrics in outcomes research: the use of instrumental variables. *Annu Rev Public Health* 1998;**19**:17-34.
27 Senn S. Change from baseline and analysis of covariance revisited. *Stat Med* 2006.
28 Lord FM. A Paradox in the Interpretation of Group Comparisons. *Psychological Bulletin* 1967;**68**:304-5.
29 Wainer H, Brown LM. Two statistical paradoxes in the interpretation of group differences: illustrated with medical school admission and licensing data *The American Statistician* 2004;**58**:117-23.
30 Frost C, Thompson SG. Correcting for regression dilution bias: comparison of methods for a single predictor variable. *Journal of the Royal Statistical Society, Series A* 2000;**163**:173-89.
31 Pettifor AE, Kleinschmidt I, Levin J, *et al.* A community-based study to examine the effect of a youth HIV prevention intervention on young people aged 15-24 in South Africa: results of the baseline survey. *Trop Med Int Health* 2005;**10**:971-80.
32 Cornfield J. Randomization by group: a formal analysis. *American Journal of Epidemiology* 1978;**108**:100-2.
33 Victora CG, Habicht JP, Bryce J. Evidence-based public health: moving beyond randomized trials. *Am J Public Health* 2004;**94**:400-5.
34 Shadish WR, Cook TD, Campbell DT. Quasi-experiments: interrupted time-series designs. *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Boston: Houghton Mifflin 2001.

35 Biglan A, Ary D, Wagenaar AC. The value of interrupted time-series experiments for community intervention research. *Prev Sci* 2000;**1**:31-49.
36 Grijalva CG, Nuorti JP, Arbogast PG, *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.
37 Clancy L, Goodman P, Sinclair H, *et al.* Effect of air-pollution control on death rates in Dublin, Ireland: an intervention study. *Lancet* 2002;**360**:1210-4.
38 Perez A, Dennis RJ, Rodriguez B, *et al.* An interrupted time series analysis of parenteral antibiotic use in Colombia. *J Clin Epidemiol* 2003;**56**:1013-20.
39 Brumback B, Ryan L, Schwartz J, *et al.* Transitional regression models, with application to environmental time series. *J Am Statist Ass* 2000;**95**:16-27.
40 Cole SR, Hernan MA, Robins JM, *et al.* Effect of highly active antiretroviral therapy on time to acquired immunodeficiency syndrome or death using marginal structural models. *Am J Epidemiol* 2003;**158**:687-94.