Using natural experiments to evaluate population health interventions


Marlborough Theatre, King’s Fund, 11–13 Cavendish Square, London, 21 January 2010

Introduction

In many circumstances, randomised controlled trials are the method of choice for evaluating interventions because randomisation, coupled with design features such as blinding and concealment of the allocation sequence, provide the most robust ways of minimising selection, allocation and other common biases. Exceptions are where randomisation is impossible, for example in the case of health protection legislation, or where the risk of bias is very low, for example where the intervention has rapid, large effects that are unlikely to reflect ‘confounding by indication’ or any other form of bias. But there are situations, such as the evaluation of service re-organisation or area-based public health interventions, in which randomisation is difficult for practical or political reasons, but where the likelihood of bias is high due to the risk of contamination, selection or small effects that take time to emerge. In such cases, observational and quasi-experimental designs are not necessarily an appropriate solution to the problems of planned experimentation, and it is important that the choice between methods is based on a full appreciation of their advantages and disadvantages.

The factors that should inform the choice have been discussed in several recent publications.\(^1-10\) There have also been a number of head-to-head comparisons of randomised and non-randomised evaluations, though such comparisons are only possible in a small (and probably biased) sample of cases, and the results are inconclusive and difficult to interpret from the point of view of choosing between alternative designs.\(^11-14\) ‘Natural experiments’ have been advocated as a solution to the problems of randomising exposure to public health interventions\(^15\,16\) but the criteria that constitute a good natural experiment are far from clear. Given the risk of bias associated with non-randomised designs, appropriate measurement of potential confounders and statistical adjustment is vital, and again, the choice of methods should be based on a full appreciation of the options. The fact that methods such as propensity scoring and other alternatives to conventional multiple regression\(^17\,18\) rarely appear in the evaluation literature suggest that they are not sufficiently widely understood.

This suggests that clear, evidence-based guidance on the circumstances in which non-randomised designs can be used reliably would be useful for researchers and commissioners of evaluations. It should help to ensure that appropriate research designs and analysis methods are chosen, and also that decisions about
implementation are based on a clear understanding of their implications for opening up or closing down evaluation options.

Aims of the workshop

The MRC Population Health Sciences Research Network, with the support of the MRC Methodology Research Panel convened the workshop to begin the process of producing such guidance by addressing questions such as: What are the criteria that define a good natural experiment? What are the circumstances in which natural experiments have been used effectively? What analysis methods are needed to obtain the most reliable estimates of effect from non-randomised evaluation designs? What are the main threats to the validity of such designs? Are there effective methods of dealing with these threats? What are the most promising approaches to improving the methodology of evaluating natural experiments?

The workshop was meant to appeal to those with an interest in the use of natural experiments to evaluate population health interventions, including producers, users, funders and publishers of research. It was hoped that, depending on the outcome of the discussions, a small working group could be convened to draft guidance which could be published as an MRC Guidance Document and an accompanying journal article, and develop proposals for further methodological research.

Proceedings

Opening the workshop, Professor Cyrus Cooper drew attention to the assumption made by some policy documents that natural experiments were a general solution to the difficulty of acquiring evidence of the effectiveness of public health interventions. For example, the 2004 Wanless report, Securing good health for the whole population, commented that, ‘the major constraint to further progress on the implementation of public health interventions is the weakness of the evidence base regarding their effectiveness and cost-effectiveness,’ and went on to suggest that, ‘current public health policy and practice, which includes a multitude of promising initiatives, should be evaluated as a series of natural experiments.’ The workshop would take a more discriminating approach, and seek to identify examples of successful natural experiments, while acknowledging the pitfalls, and areas where methods needed to be improved.

Professor Sir Michael Rutter set out the rationale for using natural experiments in health research. Understanding causality is important, and methods are needed to test causal inferences from observed associations. Statistical adjustment for confounding goes some way towards this, but is insufficient especially when confounders are imperfectly measured or unmeasured. By exploiting circumstances in which variables that normally go together are separated, natural experiments provide a range of opportunities for identifying environmental risks, and for dealing with selection bias, reverse causation and unmeasured confounders. Although randomised controlled trials are currently under-used in public health, they are often impractical, and natural experiments are a valuable alternative, although secure inferences may only be made using a variety of designs, and ingenuity is needed to identify good opportunities.
Discussion focussed on the applicability of natural experimental methods to complex public health interventions, whether they could be used to assess harm, and the importance of understanding causality as opposed to simply estimating effect sizes. In principle, natural experiments could be used to address harmful effects, but this could be difficult in practice if they took longer to emerge than the benefits of an intervention. It is always useful to understand causal mechanisms, so that interventions can be improved, and methods for doing this, such as mediation analyses, are under-used.

**Professor Barney Reeves** discussed epidemiological issues in the use of natural experiments from the point of view of incorporating evidence from such designs within systematic reviews. Although many forms of bias affect both randomised and non-randomised studies (the main exception being selection bias), the higher risk of all biases in non-randomised studies means that a cautious approach is needed. Systematic reviews of non-randomised studies should try to display evidence in an impartial way, for example using forest plots rather than relying on narrative summaries. Pooling results from studies judged to be at high risk of bias may be of limited value because the additional uncertainty from the high risk of bias undermines the extra precision from larger numbers of observations. Although attention has traditionally focussed on selection bias in exposure to the intervention, selective reporting is also a serious problem. Research on selective reporting has, to date, focussed on selection of outcomes, but the principle of selective reporting is likely to extend to selection of models when carrying out complex analyses on observational data.

Given the risks involved in using non-randomised designs, they should not be seen as a ‘quick and dirty’ option. A systematic approach is needed to weighing up the value of the evidence that can be obtained, and research questions prioritised accordingly. Possible methods for improving the quality of evidence from non-randomised studies included: prospective registers and publication of detailed protocols and a priori analysis plans to reduce selective reporting; concentration of evaluation resources on the most promising opportunities for using non-randomised designs, and involvement of methodologists in the design of such studies. Both producers and users of evidence from non-randomised studies should adopt a ‘sceptical prior’ rather than looking for evidence to support or justify a prior position. Discussion focussed on the question whether questionable evidence was always better than none. It was suggested that evidence from otherwise uncontrolled before and after studies such as speed cameras was so biased by ‘confounding by indication’ that it could only show improvement. Although it may be risky to base decisions on poor evidence, discarding the evidence in favour of expert judgement begged the question of the basis of those judgements. Developing approaches for minimising bias, and taking into account the impact of any remaining biases on effect estimates was a more promising way forward.

**Dr David Ogilvie** discussed the practicalities of conducting natural experiments, asking, first, which natural experiments should be evaluated and what questions should be asked of them, and secondly, what were the key technical questions. Policy experiments take place all the time. They may involve interventions of great public importance, such as a new urban motorway, where the costs and consequences are potentially large, but where the evidence is weak. While evaluation of such interventions is clearly desirable, questions remain about which to concentrate on, and what criteria should guide the choice, given that the most important, in terms of cost or impact, are not necessarily the most tractable from a research point of view. The most appropriate aim of the evaluation may
not be to provide an overall effect estimate, but to assess some aspect of the intervention, and to understand how effects vary by setting or population, or to identify causal mechanisms. Key technical questions concern the definition of the study population, the range of data that should be collected on exposure, outcomes, mediators and confounders, and the choice of controls. It is unlikely that an ideal design will be possible, even if all these issues are carefully considered, and even if there is some scope for the researchers to influence the way the intervention is implemented. Expectations should be realistic, and individual studies should be seen as contributing to the accumulated evidence base rather than providing the last word on effectiveness.20

In discussion it was suggested that researchers should maximise the potential for evaluation by seeking out opportunities to incorporate an element of planning within natural experiments, for example by prior investment in routine datasets, or by influencing policy makers to implement new programmes in ways that are amenable to evaluation, and research funders to act rapidly and responsively. The point was also made that some public health interventions are essentially ‘homeopathic’ – that is, they are so unlikely to have any impact that whatever the investment in the intervention, evaluation is barely worthwhile.

Professor Matt Sutton provided an overview of economists’ uses of natural experiments, and of the way they have approached some of the methodological problems. Natural experiments have been widely used in economics, for example to estimate the returns to education21,22 or the impact of labour market policies,23 because of the difficulty of conducting randomised controlled trials in these areas. A range of approaches has been developed to deal with the bias that occurs when there are observed or unobserved differences between participants and non-participants, other than exposure to the intervention, that are associated with variation in outcomes. Difference-in-difference designs compare changes in participants and non-participants in a programme, and use matching or statistical adjustment to minimise or control for observed differences between the groups. Compositional changes, differing macro trends or unobserved differences between groups may still bias estimates of programme effects.

An alternative approach, that addresses selection on unobserved differences, involves the use of an instrumental variable that affects participation in a programme but is otherwise unassociated with the outcome, and thereby mimics random assignment in a trial. A third approach is the regression discontinuity design which compares individuals who are just above and below a threshold in a continuous variable that determines eligibility to participate in a programme, and who are therefore otherwise similar. Like the instrumental variable approach it will only provide a good estimate of the effect of the programme if individuals do not react to their assignment. The plausibility of the assignment rule is a major source of debate in relation to these designs, and a number of approaches are available to test the key assumptions and assess the robustness of results.

Discussion focused on the kind on the types of policy to which these methods could be applied. Variables that provide good instruments often have very small effects. Often some incidental aspect of a policy or programme is used, so that the evaluation is of a specific exposure rather than of the intervention as a package.

In the first of the afternoon presentations of case studies, Sally Haw described the evaluation of the legislative ban on smoking in public places in Scotland. The ban was a relatively simple and well-defined intervention, but one with multiple impacts across a wide range of settings. A challenge in designing the evaluation
was that the precise details of the legislation were uncertain until shortly before enactment, which meant that the timescale for commissioning the work was very short. A range of studies was commissioned, using a variety of methods, including partly controlled before and after designs. Strengths of the evaluation included a strong body of prior evidence, which suggested that the ban might have a large, rapid impact on second-hand smoke exposure, precise measurement of exposure and confounders, a portfolio approach rather than reliance on a single design, and scope to assess alternative explanations for the observed changes. Close collaboration between researchers and policy makers greatly increases the possibility of good quality evaluation of this kind of policy change.

**Professor Ted Melhuish** described the Effective Preschool and Primary Education (EPPE) study and the National Evaluation of Sure Start (NESS). The EPPE study, a longitudinal study of children from age 3 until 11, found that children who went to the most effective preschools, had similar outcomes, in terms of attainment in reading and maths, regardless of the effectiveness of their primary schools, but there were substantial differences in outcomes between primary schools for those children who attended less effective preschools. The NESS study had three components: a cross sectional study comparing children and families in Sure Start areas shortly after implementation, with those in areas scheduled for later implementation; a study of variability in programme implementation, and a longitudinal study comparing children in Sure Start areas and non-Sure Start areas. The control areas for the longitudinal study were drawn from sampling points in the Millennium Cohort Study (MCS), using propensity score matching on a wide range of social and demographic characteristics, to select areas comparable to the Sure Start areas. The earlier study found a range of positive outcomes, but also some adverse outcomes for single mothers in the Sure Start areas. There were no adverse outcomes in the longitudinal study, and a range of positive effects, with no significant variability across different types of household. The difference between the results of the two studies may reflect the way the programme has developed in the meantime, but may also reflect differences in the methods. For example, the use of propensity scores excluded the most deprived areas from the comparison as there were almost no equally deprived areas in the MCS, though subsequent analysis suggested that the outcome models applied equally well to the excluded as to the included areas.

**Professor David Gunnell** described the way natural experiments have been used to evaluate the impact on suicide rates of events such as economic crises, changes in the availability or lethality of commonly used means of suicide, or specific interventions designed to prevent suicide. The low rate of suicide, even amongst relatively high-risk groups such as those recently discharged from a psychiatric hospital, means that randomised controlled trials of suicide prevention are impractical. Natural experiments have provided a highly effective alternative in a variety of settings, providing convincing evidence for the effects of the replacement of coal gas with natural gas in the UK, legislation to ban imports of pesticides commonly used in self-poisoning in Sri Lanka, the East Asian economic crisis, and the withdrawal of co-proxamol in the UK. The changes in some cases are so dramatic that graphical evidence is compelling, but graphical methods have been supplemented by time series and regression-based statistical approaches.

**Kenny Lawson** used the evaluation of a national programme of social housing and neighbourhood improvement in Scotland (the SHARP study) to pose the
question: what does an economic evaluation add to the evaluation of a natural experiment. The main contribution is to make the results more useful to decision-makers by including a comprehensive generic outcome measure(s) and to collect relevant costs. This enables alternative interventions to be compared in terms of their cost-effectiveness and responds to policymakers’ demands for more comprehensive evaluations that generate robust evidence of not only “what works”, but crucially whether it is “value for money” in order to prioritise future scarce resources. To generate this information economists need to be consulted at the design phase of a study.

In addition, one of the most important contributions economists can make is through prior modelling to establish the “value of the information” (VOI) to be gained from conducting the evaluation and assess whether a further study is required. This process can also focus the study design upon key areas of uncertainty and ensure there is sufficient statistical power to draw valid and generalisable inferences.

Introducing the discussion, Professor David Leon suggested that while the definition of a natural experiment is unproblematic (though a glossary would nonetheless be useful), it is less clear what kind of questions they are best suited to answer. Natural experiments have been used in three main ways in health research: to understand mechanisms by which health can be improved, a useful though not always necessary step in the development of effective interventions; to evaluate interventions intended to improve health, where they had contributed to expanding the stock of evidence, and helped defend the gains from successful interventions, such as bans on smoking in public places; and they had been used to evaluate the health impacts of interventions, such as congestion charges, that may impact on health, but are motivated primarily by other goals. Although there were good examples of the first two, there were fewer for the third. In the light of this, Professor Leon asked, what are the issues that guidance for researchers and policy makers should cover?

Issues suggested by delegates included:

A call to researchers to press policy-makers to use the best possible methods, and to advocate implementation approaches that would allow for experimentation, rather than automatically relying on observational methods.

The guidance should use successful examples of natural experimental approaches to demonstrate where they can be used appropriately, rather than focusing on the difficulties and weaknesses of the approach.

Rather than presenting a toolbox of standard techniques, the guidance should aim to help researchers recognise opportunities for natural experiments, and to apply experimental thinking and methods appropriately.

Planned and natural experiments should not be seen as entirely distinct approaches. Instead, researchers should look out for opportunities to incorporate planned experiments within larger-scale implementation of policies.

Natural experiments should be approached with the same care as trials – with close attention to issues such as statistical power, unanticipated consequences, adverse outcomes, etc., and clarity about the perspective adopted for the purposes of economic evaluation.
It is important to be clear who the guidance is intended for, and to produce different versions for different audiences. Although some of the issues may be quite technical, boxes, checklists, etc., could be used to improve accessibility to lay readers.

Improvements in the collection of data routinely, and in linking administrative datasets, and making them accessible to researchers, could greatly increase the opportunities for natural experiments.

Reporting guidelines should be included with a view to standardising the reporting of potential biases, methods used to reduce their impact, etc. Those conducting a study are best-placed to assess bias, and a structured approach to reporting would improve the prospects for informative evidence synthesis.

Closing the workshop, Peter Craig thanked speakers and delegates for their contributions. There was a consensus that guidance on the use of natural experiments would be useful and proposals for producing this would be circulated with the report of the workshop.
Annex A Programme

09.30  Coffee and registration

10.00  Introduction and aims of the workshop

10.10  Professor Sir Michael Rutter, Natural experiments – rationale and overview

10.40  Professor Barney Reeves, Epidemiological issues in the design and conduct of natural experiments

11.25  Break

11.30  Dr David Ogilvie, Practical issues in the design and conduct of natural experiments

12.15  Professor Matt Sutton, Economists’ use of natural experiments

13.00  Lunch

13.40  Case studies
   1. Ms Sally Haw: Public smoking bans
   2. Professor David Gunnell: Suicide prevention
   3. Professor Ted Melhuish: Early years studies
   4. Mr Ken Lawson: Scotland’s Housing and Regeneration Project

15.00  Break

15.15  Facilitated discussion: what are the key issues that guidance for the use of natural experiments to evaluate population health interventions should address?

16.00  Conclusion: next steps in developing guidance

16.30  Close
Annex B References


Annex C Participants

Dr Suzanne Audrey  Department of Social Medicine, University of Bristol
Dr Janis Baird  MRC Epidemiology Resource Centre, University of Southampton
Clare Beeston  NHS Health Scotland
Christina Black  MRC Epidemiology Resource Centre, University of Southampton
Professor Lyndal Bond  MRC/CSO Social and Public Health Sciences Unit, University of Glasgow
Zoltan Bozoky  Policy Research Programme, Department of Health
Professor Iain Buchan  Health Methodology Research Group, University of Manchester
Dr Bryony Butland  Cross Government Obesity Team, Department of Health
Dr Dexter Canoy  School of Community Based Medicine, University of Manchester
Professor Roy Carr Hill  Centre for Health Economics, University of York
Catherine Chittleborough  Department of Social Medicine, University of Bristol
Dr Andrew Cook  Public Health Research Programme, NIHR Evaluation, Trials and Studies Co-ordinating Centre
Sarah Cook  Department of Epidemiology and Population Health, London School of Hygiene and Tropical Medicine
Professor Cyrus Cooper  MRC Epidemiology Resource Centre, University of Southampton
Neil Craig  NHS Health Scotland
Dr Peter Craig  MRC Population Health Sciences Research Network
Dr Steve Cummins  Department of Geography, Queen Mary, University of London
Professor Linda Davies  Health Methodology Research Group, University of Manchester
Neil Davies  Department of Social Medicine, University of Bristol
Dr Monica Dent  National Obesity Observatory and Public Health Resource Unit
Dr Matt Egan  MRC/CSO Social and Public Health Sciences Unit, University of Glasgow
Ibidun Fakoya  Research Department of Infection & Population Health, University College London
Dr Elizabeth Fenwick  Public Health and Health Policy, University of Glasgow
Dr Bruna Galobardes  Department of Social Medicine, University of Bristol
Dr Darya Gaysina  MRC Unit for Lifelong Health and Ageing, University College London
Dr Simon Griffin  MRC Epidemiology Unit, Institute of Metabolic Science
Professor David Gunnell  Department of Social Medicine, University of Bristol
Sally Haw  Scottish Collaboration for Public Health Research and Policy
Neil Hawkins  Oxford Outcomes
Dr Andrew Hayward  UCL Centre for Infectious Disease Epidemiology, University College London
Dr Julian Higgins  MRC Biostatistics Unit, University of Cambridge
Dr Melvyn Hillsdon  School of Sport and Health Sciences, University of Exeter
Professor Frank Kee  UKCRC Centre of Excellence in Public Health (NI), Queen’s University of Belfast
Dr Andrew Jones  School of Environmental Sciences, University of East Anglia
Professor Catherine Law  NIHR Public Health Research Programme, NIHR Evaluation, Trials and Studies Co-ordinating Centre
Kenny Lawson  Public Health and Policy, University of Glasgow
Dr Frank van Lenthe  Department of Public Health, Erasmus Medical Centre, Rotterdam
Professor David Leon  Department of Epidemiology and Population Health, London School of Hygiene and Tropical Medicine
Professor Alastair Leyland  MRC/CSO Social and Public Health Sciences Unit, University of Glasgow
Professor Anne Ludbrook  Health Economics Research Unit, University of Aberdeen
Professor Ronan Lyons  Centre for Health Information, Research and Evaluation, Swansea University
Professor Sally Macintyre  MRC/CSO Social and Public Health Sciences Unit, University of Glasgow
Professor Edward Melhuish  Birkbeck College, University of London
Dr Tim Millar  National Drug Evidence Centre, University of Manchester
Graham Moore  DECIPHer, UKCRC Public Health Centre of Excellence, Cardiff University
Professor Laurence Moore  DECIPHer, UKCRC Public Health Centre of Excellence, Cardiff University
Simon Murphy  DECIPHer, UKCRC Public Health Centre of Excellence, Cardiff University
Professor Jon Nicholl  Medical Care Research Unit, University of Sheffield
Dr David Ogilvie  MRC Epidemiology Unit, Institute of Metabolic Science
Lesley Owen  Centre for Public Health Excellence, National Institute of Health and Clinical Excellence
Dr Tom Palmer  MRC Centre for Causal Analyses in Translational Epidemiology, University of Bristol
Dr Craig Ramsay  Health Services Research Unit, University of Aberdeen
Professor Barney Reeves  Clinical Trials and Evaluation Unit, University of Bristol
Dr Hannah Rothstein  MRC Biostatistics Unit, University of Cambridge
Dr Harry Rutter  National Obesity Observatory, University of Oxford
Professor Sir Michael Rutter  Social, Genetic and Developmental Psychiatry Centre, Institute of Psychiatry
Professor Roy Sainsbury  Social Policy Research Unit, University of York
Rita Santos  Centre for Health Economics, University of York
Dr Linda Sheppard  Centre for Public Health Excellence, National Institute of Health and Clinical Excellence
Kathryn Skivington  MRC/CSO Social and Public Health Sciences Unit, University of Glasgow
Dr Pam Sonnenberg  Research Department of Infection & Population Health, University College London
Rebecca Steinbach  Public and Environmental Health Research Unit, London School of Hygiene and Tropical Medicine
Professor Matthew Sutton  Health Methodology Research Group, University of Manchester
Professor Simon Thompson  MRC Biostatistics Unit, University of Cambridge
Professor Catharine Ward Thomson  Edinburgh College of Art, University of Edinburgh
Kathryn Wheeler  Department of Geography, Queen Mary, University of London
Professor Martin White  UKCRC Centre for Translational Research in Public Health, Newcastle University
Lin Yang  MRC Epidemiology Unit, Institute of Metabolic Science