Using natural experiments to evaluate population health interventions:

guidance for producers and users of evidence

Prepared on behalf of the Medical Research Council by:

Peter Craig, Programme Manager, MRC Population Health Sciences Research Network

Cyrus Cooper, Director, MRC Lifecourse Epidemiology Unit

David Gunnell, Professor of Epidemiology, Department of Social Medicine, University of Bristol

Sally Haw, Principal Public Health Adviser and Associate Director, Scottish Collaboration for Public Health Research and Policy

Kenny Lawson, Research Fellow in Health Economics, Centre for Population and Health Sciences, University of Glasgow

Sally Macintyre, Director, MRC|CSO Social and Public Health Sciences Unit

David Ogilvie, Clinical Investigator Scientist and Honorary Consultant in Public Health Medicine, MRC Epidemiology Unit, University of Cambridge

Mark Petticrew, Professor of Public Health Evaluation, Department of Social and Environmental Medicine, London School of Hygiene and Tropical Medicine

Barney Reeves, Professorial Fellow in Health Services Research and co-Head, Clinical Trials and Evaluation Unit, University of Bristol

Matt Sutton, Professor of Health Economics, Health Methodology Research Group, School of Community-based Medicine, University of Manchester

Simon Thompson, Director, MRC Biostatistics Unit

www.mrc.ac.uk/naturalexperimentsguidance

Table of Contents

Acknowledgements ... 3

Summary ... 4

Aim of guidance ... 4
What are natural experiments? ... 4
When should they be used? ... 4
How can design, analysis and reporting be improved? ... 4
Conclusions ... 5

1. Introduction ... 6

1.1 What are natural experiments? ... 6
1.2. Why are natural experiments important? ... 7

2. How have natural experiments been used in the population health sciences? ... 8

2.1. Understanding mechanisms ... 9
2.2. Evaluating interventions to protect or improve health ... 9
2.3. Evaluating interventions where health impacts are by-products ... 11

3. Improving the use of natural experiments ... 13

3.1. Choosing when to use a natural experimental approach ... 13
3.2. Natural and planned experiments ... 14
3.3. Assessing value for money ... 15

4. Improving the design, analysis and reporting of natural experiments ... 17

4.1. Design ... 17
4.2. Analysis ... 18
4.2.1. Selection on observables ... 18
4.2.2. Selection on unobservables ... 19
4.3. Strengthening causal inference ... 21
4.4. Reporting natural experiments ... 22
4.5. Systematic reviews of natural experiments ... 23

5. Conclusion ... 24

Bibliography ... 25

Annex 1: Alternative definitions of natural experiments ... 29

Acknowledgements

The production of this guidance has been supported by the MRC Population Health Sciences Research Network and the MRC Methodology Research Panel. We thank the participants at workshops at the King's Fund in January 2010, and the Society for Social Medicine Annual Scientific Meeting in September 2010. We also thank the following for helpful comments on drafts of the document: Amina Aitsi-Selmi, University College London, Catherine Chittleborough, University of Bristol, Andrew Cook, Ruairidh Milne and James Raftery, National Evaluation Trials and Studies Co-ordinating Centre, Neil Craig, NHS Health Scotland, Matt Egan, MRC|CSO Social and Public Health Sciences Unit, Catherine Law, University College London and NIHR Public Health Research Programme, Ann Prentice, MRC Human Nutrition Research, Nick Wareham, MRC Epidemiology Unit, Martin White, FUSE, University of Newcastle. The authors take full responsibility for any remaining errors.

Summary

Aim of guidance

The aim of this guidance is to help producers, users, funders and publishers of evidence understand how and when 'natural experiments' can be used to good effect. While there is growing interest in using natural experiments to evaluate large-scale interventions to improve population health, the relevant methodological literature is dispersed across disciplines and there is a lack of general guidance either on the range of approaches available, or on the circumstances in which they are likely to be useful. We aim to fill this gap.

What are natural experiments?

By natural experiments, we mean events, interventions or policies which are not under the control of researchers, but which are amenable to research which uses the variation in exposure that they generate to analyse their impact. By natural experimental studies, we mean the methodological approaches to evaluating the impact on health or other outcomes of such events. The key features of these definitions are that (1) the intervention is not undertaken for the purposes of research, and (2) the variation in exposure and outcomes is analysed using methods that attempt to make causal inferences. Classic examples include the effect of famine on the subsequent health of children exposed in utero, or the effects of clean air legislation, indoor smoking bans, and changes in taxation of alcohol and tobacco. The event of interest could involve the introduction of new legislation (eg to ban the use of polluting fuels), withdrawal or amendment of an existing policy (eg the abolition in 1999 of GP fundholding), or changes in the level of an intervention or service (eg an increase in alcohol taxation or a change in the nutritional requirements of school meals); it could also be an event far removed from health policy, such as an economic downturn or upturn, or an agreement on international trade.

When should they be used?

Natural experimental approaches widen the range of events, policies or interventions that can usefully be evaluated beyond those that were designed for research purposes, but they are not always suitable. The case for a natural experimental study is strongest when: there is scientific uncertainty about the size or nature of the effects of the intervention; for practical, political or ethical reasons, the intervention cannot be introduced as a true experiment, for example in a randomised controlled trial; it is possible to obtain the relevant data from an appropriate study population, in which exposed and unexposed groups — or groups with different levels of exposure — can be compared; and the intervention or the principles behind it have the potential for replication, scalability or generalisability. A 'value of information' analysis can help to make a convincing case for a natural experimental study, and economic evaluation may make the results more useful for decision-makers.

How can design, analysis and reporting be improved?

Randomised controlled trials and natural experimental studies are both subject to similar threats to validity. The key difference is that while randomised trials have a very general method of minimising the bias caused by selective exposure to the experimental intervention, in the case of non-randomised studies there is a range of partial solutions. If an intervention is expected to have a very large or rapid impact, a simple design may appear to be adequate. Natural experiments can be used to study more subtle effects, so long as a suitable source of variation in exposure can be found, but the design and analysis become more challenging. Whatever the expected effect size, care should be taken to minimise bias. Combinations of methods, testing and sensitivity analysis should be used to provide additional checks on the plausibility of causal inferences.

Transparent reporting of natural experimental studies is also vital. Established guidelines such as STROBE or TREND should be followed, with particular attention to: clearly identifying the approach as a study of a natural experiment; providing a clear description of the intervention and the assignment process; and explicitly stating the methods used to estimate impact. Procedures used to reduce bias should be discussed in a detailed and balanced way. Ideally, qualitative judgements about the risk of bias, and how well it has been dealt with, should be supplemented by a quantitative assessment. If the study used multiple methods, variation in the estimates should be highlighted. The context within which the intervention was implemented should be described as this may affect interpretation and help users assess the generalisability of the findings. Wherever possible, the results should be compared with those of other evaluations of similar interventions, paying attention to any associations between effect sizes and variations in evaluation methods and intervention design, content and context.

Conclusions

Natural experimental approaches have a great deal to offer. We have learnt much of epidemiological and policy value from studies of, for example, the replacement of coal gas with natural gas in Britain and the implementation of indoor smoking bans in many countries around the world. Natural experimental approaches are not restricted to situations where the effects of an intervention are large or rapid; they can be used to detect more subtle effects where there is a transparent exogenous source of variation. Even so, it would be unwise to assume that such approaches are suitable for evaluating particular policies or interventions without very detailed consideration of the kind of opportunities the intervention will generate. Optimism about the use of natural experiments should not be a pretext for discounting the option of conducting planned experiments, where these are possible and would be more robust. Randomised trials will often be the only way to obtain reliable estimates of effect, and some research questions may be genuinely intractable. Research effort should be focused on addressing important and answerable questions, taking a pragmatic approach based on combinations of methods, plus explicit recognition and careful testing of assumptions. Priorities for the future are to build up experience of promising but lesser used methods, and to improve the infrastructure that enables opportunities presented by natural experiments to be seized, including good routine data from population surveys and administrative sources, good working relationships between researchers and policy makers, and flexible forms of research funding.

1. Introduction

There is growing interest in the use of natural experiments – events that are not planned for the purposes of research - to evaluate population health interventions. The 2004 Wanless report, ***Securing good health for the whole population***, suggested that 'current public health policy and practice, which includes a multitude of promising initiatives, should be evaluated as a series of natural experiments.'[1] The 2007 Foresight report on obesity[2] called for more support for the evaluation of natural experiments 'relative to highly controlled experimental paradigms,' a call echoed the following year by the Department of Health in its obesity research and surveillance plan, ***Healthy weight, healthy lives***.[3] A similar growth of interest is evident in other areas where planned experimentation is difficult, such as political science,[4] economics[5] and history.[6] Natural experiments have been used to good effect to answer questions as diverse as whether the cost of voting affects turnout,[7] whether class size affects educational outcomes[8] and why some post-colonial countries prosper and others do not.[9] Within epidemiology there is a tradition of using major external shocks such as famines[10] and other unusual situations[11] to study the effects of environmental exposures on human development and susceptibility to disease. John Snow's investigation of the causes of a cholera outbreak in mid nineteenth century London is often cited[7 12-16] as an early, perhaps the first, example of the study of a natural experiment.

A difficulty in applying similar methods to the evaluation of population health interventions is that very often the change in exposure is much less extreme. Some interventions, such as public smoking bans[17] or legislation to control imports of pesticides frequently used for suicide,[18] do seek to remove or drastically reduce exposure to some risk factor, with immediate consequences for health. Many others, such as strategies to improve diet, encourage physical activity or reduce harmful drinking, seek more subtle effects that may take some time to emerge. Although natural experiments have certain advantages over planned experiments, it is impossible in the former to manipulate exposure to the intervention. In consequence, natural experimental studies are susceptible to bias to a much greater extent. For these reasons, it is important to be able to distinguish situations in which natural experimental approaches are likely to be informative, from those in which some form of fully experimental method such as a randomised controlled trial (RCT) is needed, and from those in which the research questions are genuinely intractable.

This guidance is intended to help researchers and users, funders and publishers of research evidence make the best use of natural experimental approaches to evaluating population health interventions. We hope that researchers will find the guidance useful in planning and designing evaluations of public health interventions. We hope that journal editors and reviewers will find it helpful in assessing the quality of studies that use observational data to evaluate interventions. And we hope that it will help policy-makers and others to recognise both the strengths and the limitations of a natural experimental approach, and to use it where appropriate to complement a fully experimental approach.

Given the variety of natural experiments, it would not be practical or useful to provide a template for producers and users of research to follow. Instead, we provide a range of case studies to draw attention to successful examples and to illustrate the range of methods available and their strengths and weaknesses. We begin by reviewing the varying uses of the term 'natural experiment' and considering why natural experiments offer important opportunities for public health research. In Section 2 we present examples of the main uses of natural experiments in public health: to understand the causes of health and disease, to evaluate health interventions, and to evaluate non-health interventions with potentially important consequences for health. In Section 3 we review ways of improving the use of natural experiments, with a focus on the circumstances in which a natural experimental approach is likely to be useful and how the information can be made useful for decision-makers. In Section 4 we look at ways of improving the design, analysis and reporting of natural experimental studies. In conclusion, we draw together the main messages, and suggest some directions for the future development of natural experimental approaches as tools for public health research. While we are aiming for a broad readership with varying levels of methodological interest and experience, some of the material — especially the discussion of methods of analysis in Section 4.2 — is aimed more at researchers looking to apply those methods, whereas most of the rest of the document is aimed at a more general readership. The guidance is illustrated with examples and case studies throughout.

1.1 What are natural experiments?

The term 'natural experiment' lacks an exact definition. It tends to be used interchangeably to refer either to the methodological approach of using unplanned or uncontrolled events as a source of variation, or to the events themselves. At its broadest, it has been used to distinguish the 'comparative method', i.e. detailed comparisons of contrasting cases, from single case studies.[6] At the other extreme, it is used to refer to studies in which there is no manipulation of exposure, but the assignment of subjects is 'as if' random.[7] (Annex 1) Most definitions lie between these extremes, and characterise natural experimental studies as those which exploit natural or unplanned variation in exposure, i.e. variation that is not manipulated for the purposes of research, using a combination of design and analytical features that are meant to allow causal inferences to be drawn. Some authors distinguish such studies from straightforward observational studies where no intervention takes place, and from the large (but also imprecisely defined) class of planned but non-randomised experiments sometimes referred to as quasi-experiments.[19]

For the purposes of this guidance, we use the term natural experiment to refer to the event of interest. We use 'natural experimental study' to refer to ways of evaluating interventions using unplanned variation in exposure (in the sense given above) to analyse impact. The key features of these definitions are that (1) the intervention is not undertaken for the purposes of research, and (2) the variation in exposure and outcomes is analysed using methods that attempt to make causal inferences. In other words, natural experimental studies involve the application of experimental thinking to non-experimental situations. Outside of an RCT it is rare for variation in exposure to an intervention to be random, so special care is needed in the design, reporting and interpretation of evidence from natural experimental studies, and causal inferences must be drawn cautiously.

1.2. Why are natural experiments important?

Interest in alternatives to RCTs has been fuelled by policy and research interest in population-level environmental and non-health sector interventions to improve health.[20 21] Such interventions may be intrinsically difficult to manipulate experimentally, as in the case of national legislation to improve air quality or major changes in transport infrastructure.[22] It may be unethical to manipulate exposure in order to study effects on health if the intervention has other known benefits, if it has been shown to work in other settings, or if its main purpose is to achieve non-health outcomes.[23] Interventions may be implemented in ways that make a planned experiment difficult or impossible, with short timescales or extreme variability in implementation.[24] A randomised trial may be ethically sound and practically feasible but politically unwelcome, regardless of ethical or practical considerations.[25]

Natural experimental approaches are important for two reasons: (1) they widen the range of interventions that can usefully be evaluated beyond those that are amenable to planned experimentation; and (2) they encourage a rigorous and imaginative approach to the use of observational data to evaluate interventions that should allow stronger conclusions about impact. However, it is misleading to assume that whenever a planned experiment is impossible, there is a natural experimental study waiting to happen. Only a small proportion of the 'multitude of promising initiatives'[1] are likely to yield good natural experimental studies. Care, ingenuity and a watchful eye for good opportunities will be required to realise their potential.

2. How have natural experiments been used in the population health sciences? – some examples

Natural experimental studies have been extensively used to investigate environmental causes of disease, by exploiting naturally occurring, or at least unplanned, variation in exposure.[26 27] They include a range of approaches designed to discriminate between genetic and environmental causes of disease, such as twin, adoption and migration studies; 'instrumental variable' approaches that use genetic and phenotypic variants that are associated with the exposure but are not otherwise associated with the outcome of interest or with other risk factors; and designs that seek to eliminate bias caused by selective exposure to risk by examining events that result in the exposure or protection of a whole population (Section 2.1).

Natural experimental studies have also been used to evaluate interventions, including those whose primary aim is to improve or protect health (Section 2.2), and those where health impacts are secondary to some other purpose (Section 2.3). Table 1 summarises key examples, and more detailed case studies are provided in Boxes 1-7. Although there is no difference in principle between health and non-health interventions, there are important practical differences. In particular, health effects that are essentially by-products may be small or take a long time to emerge, as may some intended effects, and it may be difficult to obtain support for evaluation research that does not focus on the main aim of the intervention.[28]

Table 1: Examples of natural experimental studies to evaluate population health interventions

Population	Intervention	Comparator	Outcome	Analysis method	Results	Reference
Sri Lanka - whole population	Legal restriction on pesticide imports	Suicide rates pre-ban; method specific suicide rates; non-suicide mortality	Mortality from suicide by self-poisoning with pesticide	Graphical analysis of trends	Import bans reduced method-specific and overall suicide mortality	Gunnell et al 2007[18]
UK - 12-17 year olds	Restriction on prescribing SSRIs to people aged <18	Suicidal behaviour pre-restriction	Hospitalisations for self-harm; suicide mortality	Joinpoint regression	Restriction on prescribing of SSRIs was not associated with changes in suicidal behaviour	Wheeler et al 2008[37]
Finland - men and women aged >15	Reduction in alcohol taxes	Alcohol related mortality before the tax change	Alcohol-related mortality	Poisson regression to obtain relative rates; time-series analysis	Alcohol-related mortality increased, especially among the unemployed	Herttua et al 2008,[51] Herttua et al 2009[52]
Hong Kong - whole population	Legislation to restrict sulphur content of fuel	Mortality pre-restriction, overall and by district	All-cause and cause specific mortality	Poisson regression of change in seasonally adjusted death rates	Cardiovascular, respiratory and overall mortality fell post-restriction; decline greater in districts with larger falls in SO2 concentrations	Hedley et al 2002[43]
Dublin – whole population	Ban on coal sales	Mortality pre-ban, and in the rest of Ireland	Non-trauma, respiratory and cardiovascular mortality	Interrupted time-series	Non-trauma, respiratory and cardiovascular death rates fell post-ban	Clancy et al 2002[42]
Scotland – patients admitted to 9 hospitals	Legislative ban on smoking in public places	Hospitalisations pre-ban and in England	Hospitalisations for acute coronary syndrome	Comparison of numbers of admissions pre and post ban	Admissions for acute coronary syndrome fell among both smokers and non-smokers post-ban	Pell et al 2008[71]
England – patients aged >17	Legislative ban on smoking in public places	Hospitalisations pre-ban	Emergency admissions for myocardial infarction	Interrupted time-series	Small but significant fall in emergency admissions in the first year post-ban	Sims et al 2010[121]
India – pregnant women	Cash incentives to use a health facility to give birth	Districts with low rates of take up; births to women not receiving payments	Use of health facilities; infant and maternal mortality	Matched and unmatched comparisons of recipient and non-recipient births; difference-in-differences analysis of district level usage and mortality rates	Higher rates of take-up of the incentives were associated with higher proportions of births within health care facilities; use of health care facilities to give birth was associated with fewer perinatal and neonatal deaths. There was a non-significant reduction in maternal deaths	Lim et al 2010[97]
England – general practitioners	Abolition of GP fundholding	Non-fundholding practices; pre-abolition admission rates	Referral for elective and emergency admissions	Difference-in-difference analysis of referrals from fundholders and non-fundholders	Fundholders had lower rates of elective referral while fundholding was in operation and their rates of referral increased more than those of non-fundholders following abolition. There was no difference in emergency admissions pre or post abolition.	Dusheiko et al 2003[100]

USA – low income families with children aged 3-5	Headstart - help with parenting, nutrition, health and social services and schooling	US counties with poverty levels above the cutoff used to allocate help with accessing Headstart funding	Mortality from causes of death amenable to Headstart services	Regression discontinuity design, comparing regressions of mortality on poverty for counties above and below the cutoff	For Headstart-related causes, there was a discontinuity in mortality rates at the cutoff, but no difference in deaths from other causes or among children too old to qualify for Headstart services	Ludwig and Miller 2007[107]
USA – patients admitted to hospital with acute myocardial infarction	Invasive cardiac treatment (catherisation followed by revascularisation)	Non-invasive cardiac treatment	Long term (i.e. 7-year) mortality	Instrumental variable analysis using regional catherisation rates as the instrument	Cardiac catherisation was associated with lower mortality; instrumental variable analyses produced smaller estimates of effect than analyses (multivariate risk adjustment, propensity score-based methods) that only adjust for observed prognostic factors	Stukel et al 2007[107]

2.1. Understanding mechanisms

Risk factors for poor health frequently cluster together, making causal interpretation difficult. For example, poverty is strongly associated with mental health problems, but is poor mental health a consequence of social and economic adversity, or does the link between the two reflect a tendency for people with mental health problems to drift into poverty? Although there have been trials in which incomes were manipulated, those studies focused on labour market rather than mental health outcomes, and have fallen out of favour.[29] Instead, much work on the early origins of mental health problems has relied on natural experiments (Box 1).[30]

Box 1: The Great Smoky Mountains Study of poverty and psychopathology

This natural experimental study took advantage of the combination of a longitudinal study of the development of psychiatric disorder, which began in 1993, and the subsequent establishment of a casino on an American Indian reservation within the area covered by the study. Approximately one quarter of the children in the study belonged to families that received a royalty income from the casino's profits from 1996 onwards. Costello et al.[31] used data from the survey to classify families as persistently poor, ex-poor or never poor in order to assess the effect on children's mental health of an increase in income sufficient to lift them out of poverty.

In the children whose families moved out of poverty, the overall frequency of psychiatric symptoms fell to levels similar to those of the never poor children. It remained high in the children whose families remained poor. The findings suggest that social causation, as opposed to selection, explains the link between child poverty and psychopathology. All the Indian families received the casino income, largely breaking the link between movement out of poverty and family characteristics that might influence behavioural symptoms in the children. There was a similar pattern in the non-Indian children, suggesting that the results are generalisable, although by themselves the results in the non-Indian children would be less convincing because of the risk of confounding by other factors associated with increasing income. As the authors point out, the study therefore provides 'a fairly clean test of competing theories.'

It might seem like an amazing stroke of good fortune for a natural experiment to occur in the middle of an ongoing longitudinal study. But the researchers were able to capitalise on this 'lucky' event thanks to the rigorous design of the study, with a high response rate, low attrition and careful, repeated measures of exposure and outcome. A further follow-up in 2006, when the youngest cohort members were aged 21, showed that the protective effects of the extra income persisted into early adulthood.[32]

2.2. Evaluating interventions to protect or improve health

Suicide is rare in the general population, occurring at a rate of about 1/10,000 per annum. Even in high risk populations, such as people treated with antidepressants, the rate is only around 1/1000. Clinical trials would therefore have to be enormous to have adequate power to detect even large preventive effects.[33] There have been calls for very large, simple trials in this and other areas,[34] but in the meantime, natural experiments have been used effectively to assess the impact of measures to restrict access to commonly used means of suicide (Box 2). These findings have made restriction of suicide methods a common priority of suicide prevention strategies in the UK and worldwide.[35]

Box 2:
Suicide prevention: evidence from natural experiments

In Sri Lanka, the introduction of agricultural pesticides was followed by an epidemic rise in their use for self-poisoning in the 1960s and 1970s (Figure 1).[18]

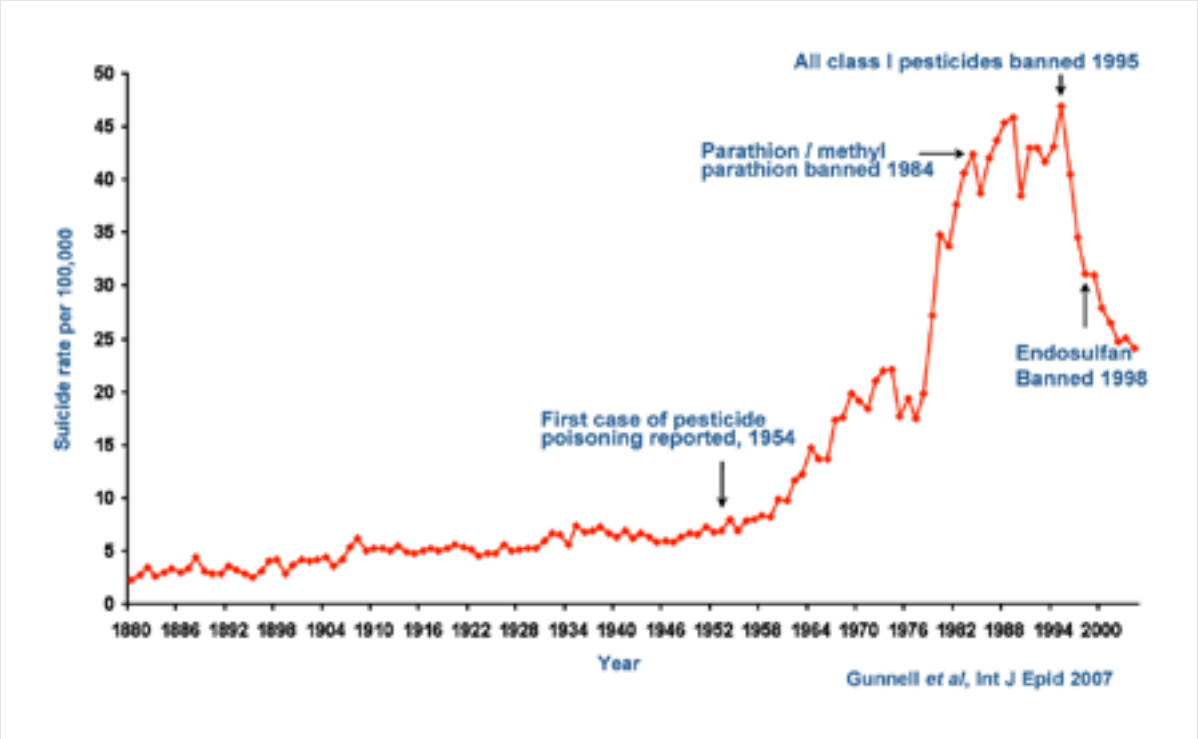

Fig.1 Suicide rate in Sri Lanka 1880-2005

The high case fatality associated with ingestion of some pesticides appears to have contributed to the four-fold rise in suicide in Sri Lanka over this period. By the 1980s it is estimated that over two thirds of Sri Lanka's suicides were due to pesticide self-poisoning. Subsequent staged bans on some of the most toxic products were associated with a halving in completed suicides despite year on year rises in the incidence of pesticide self-poisoning. Pesticide poisoning remains a major health problem in many parts of rural Asia where bans have not been implemented.

Findings from other studies demonstrating the impact of the changing availability of commonly used methods of suicide on suicide rates have been used to support regulatory action by the UK's Medicines and Healthcare products Regulatory Agency (MHRA) to restrict sales of paracetamol (in 1998) and withdraw the marketing authorisation for co-proxamol (2007). Early analysis of the impact of co-proxamol regulation demonstrates a reduction in method specific suicide rates with no evidence of a compensatory rise in the use of other medicines for fatal self-poisoning.[36] Similar designs have been used to assess the impact of regulatory restriction of antidepressant prescribing in children (Wheeler et al 2008).[37]

Where interventions consist of national regulatory activity, natural experimental designs are a key approach to evaluating their health impact. In the suicide prevention literature, methodological approaches used to analyse such interventions range from simple graphical assessments of trends,[38] to more complex methods such as interrupted time-series,[36] joinpoint regression[37] and random effects regression models[39] to assess the impact of regulatory changes across a range of countries. Such methods have recently been used to investigate the impact of periods of economic and social upheaval on suicide rates[40] and to clarify the effect of unemployment and economic crisis.[41]

Studies of legislation to reduce air pollution provide another example where a natural experimental approach has produced clear cut evidence of health impacts.[42 43] These studies have benefited from the availability of high quality, routinely collected data on exposures, potential confounders and outcomes (including outcomes that would not be expected to change) and substantial, rapid changes in exposure across a whole population, reducing the risk of selective exposure or of confounding by secular trends, and increasing the confidence with which changes in outcomes can be attributed to the intervention.

2.3. Evaluating interventions where health impacts are by-products

Natural experiments have also been used to evaluate health impacts that occur as a by-product of interventions primarily intended to influence non-health outcomes. Again, suicide provides an important example. In the 1950s and 60s domestic gas poisoning was the most frequent method of suicide in England and Wales. During the 1960s natural gas, which contains very little carbon monoxide, gradually replaced coal gas, which had a high carbon monoxide content. This led to a reduction in domestic gas suicide rates and, as there was little immediate substitution to other methods, overall suicide rates also declined and many deaths were prevented (Figure 2).[38 44]

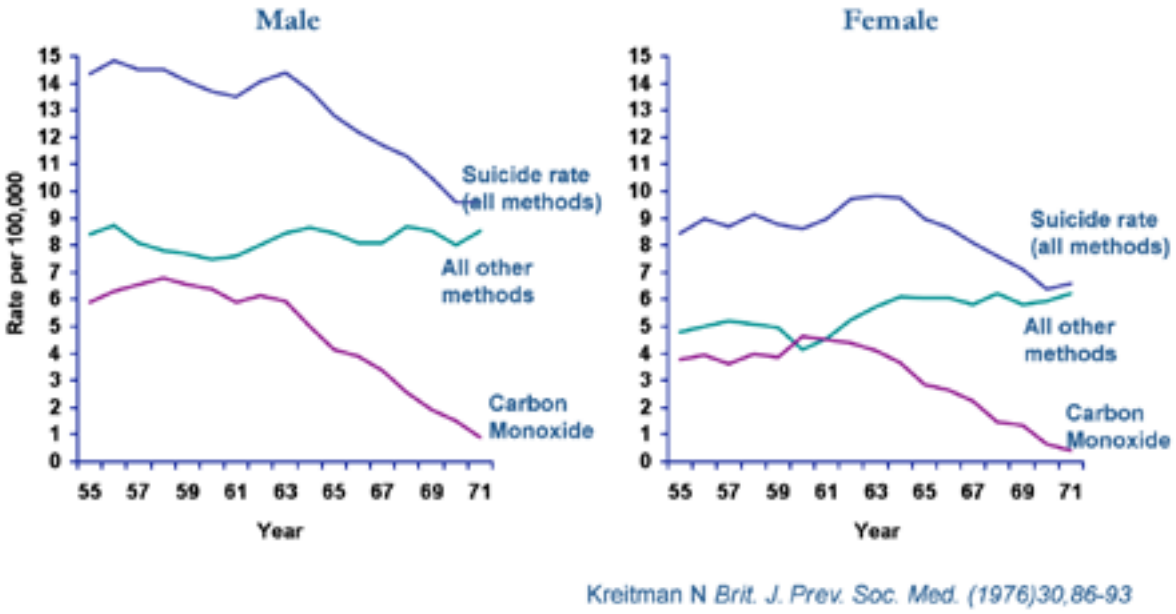

Fig.2 Sex-specific suicide rates by mode of death, England and Wales 1955-1971

Another area where natural experiments have been used extensively is in estimating the effect of changes in price, often as a result of changes in taxation, on alcohol consumption and related harms (Box 3).[45 46] In the UK, an increase in the affordability of alcohol over the past two decades has been associated with a marked increase in alcohol-related deaths, which contrasts with a fall in most Western European countries.[47] Reversing the upward trend in alcohol-related harm is now a public health priority, although the best way to achieve it, and whether price is the right mechanism, remains hotly debated, despite extensive evidence that alcohol consumption and harm respond to changes in price.[48-50]

Box 3: Alcohol prices and alcohol-related mortality

In Finland a sharp reduction in the price of alcohol in 2004 provided a natural experiment that allowed researchers to assess the relationship between a fall in price and alcohol-related mortality.[51] The deregulation of import quotas by the European Union, followed by the accession of Estonia a few months later, prompted the Finnish Government to cut taxes on alcohol by an average of one third, to prevent the country from being flooded by cheap imported alcohol. This led to a sharp reduction in the price of most alcoholic drinks, and alcohol consumption rose by an estimated 12% over the next two years.

Herttua and colleagues linked employment and death register data to compare alcohol-related mortality in 2001-3 and 2004-5 and to see whether any change varied by social position. They found that alcohol related mortality rates increased by 16% in men (an extra 22 deaths per 100,000 person-years) and 31% in women (8 extra deaths per 100,000 person-years). Subsequent analyses of survey data suggested that the patterning of the increase in mortality by age and social position matched that of the increase in consumption.[52 53]

The ability to link high-quality mortality and socio-economic data, coupled with a substantial and rapid price change, enabled the researchers to assess both the population impact of the change, and its effect on socio-economic inequalities in alcohol-related mortality. Falling prices increased alcohol related mortality, with a disproportionate impact on the unemployed.

3. Improving the use of natural experiments

The examples in Section 2 were chosen to illustrate the circumstances in which natural experiments can be used to obtain convincing evidence of causal effects or of the impacts of interventions. Perfect instances of these circumstances rarely occur in practice, and difficult choices have to be made about when to adopt a natural experimental approach and how best to exploit the opportunities that do occur. The response to recent calls to make better scientific use of opportunities to learn from natural experiments needs to be made in the context of constraints on the funding available for evaluation. The aim of this section is to provide guidance about how scarce evaluative resources may best be deployed in this area.

3.1. Choosing when to use a natural experimental approach

The fact that a natural experiment is taking place does not in itself constitute grounds for investing in a scientific evaluation. It is not necessary or feasible to evaluate every natural experiment, or every element of a given natural experiment. In practice, natural experiments form a spectrum, and the opportunities they provide for research depend on a range of factors including the size of the population affected, the size and timing of impacts, the processes generating variation in exposure, and the practicalities of data gathering. Natural experimental studies should only be attempted when exposed and unexposed populations (or groups subject to varying levels of exposure) can be compared, using samples large enough to detect the expected effects, and when accurate data can be obtained on exposures, outcomes and potential confounders. These design issues are dealt with in Section 4 below. In addition, a scientific evaluation should only be undertaken when the scientific rationale for a study can be clearly articulated. The case is strongest when as many as possible of the following conditions are satisfied. If few or none can be met, the funders or providers of a given intervention may still wish to conduct an evaluation for their own purposes, but those interests should not be confused with scientific priorities.

1. There is a reasonable expectation that the intervention will have a significant health impact, but scientific uncertainty about the size or nature of the effects. Just as some clinical research funders will not fund a randomised controlled trial unless a recent systematic review has shown the need for a new trial, any proposal for a study of a natural experiment should be similarly justified in terms of specifying the general research questions to which it will contribute new knowledge. These may be identified from systematic reviews, the research recommendations of National Institute for Health and Clinical Excellence (NICE) or other guidance, scoping reviews, expert consensus exercises and similar sources. Even when the aggregate effect of a type of intervention has been well established, important uncertainties may remain about its effects in different settings, its effects in different social groups (and therefore about its impact on health inequalities), or its wider impacts (for example the health impacts of an intervention mainly intended to reduce crime). Some natural experiments — particularly those involving simple interventions and large effects observed in multiple study populations — may be sufficient to provide a definitive answer to key research and policy questions. In many cases, however, single natural experiments in policy areas related to the wider determinants of health are unlikely to provide definitive, unambiguous scientific guidance for policy; rather, they should be regarded as contributing to a growing body of evidence for subsequent cumulation in systematic reviews and other forms of evidence synthesis.[54]

2. A natural experimental study is the most appropriate method for studying a given type of intervention. This will often depend on showing that it is impractical, unethical or politically unfeasible for the intervention to be introduced as a true experiment, at least for the foreseeable future. In most cases, a well designed and planned true experiment is likely to contribute more to the overall scientific evidence in a given area than is a hastily executed natural experiment, even if the latter type of study could be completed more quickly and would satisfy the needs of certain stakeholders. However, RCTs are clearly not possible in some situations, such as studying the impacts of interventions that are highly specific to their context (such as urban motorway construction) or otherwise unique in time and place (such as major sporting events), and it is important to avoid an 'evaluative bias' whereby evidence is mainly gathered about certain types of intervention (those amenable to an RCT) and not about others.[55]

3. It is possible to obtain the relevant data from an appropriate study population, comprising groups with different levels of exposure to the intervention. This is likely to depend on showing either that suitable data are already routinely available, or that there is enough time to specify a suitable study population and collect baseline data before the intervention is introduced. Routinely collected data have often been used successfully for the efficient evaluation of natural experiments, for example by tracking changes in the incidence of self-harm and suicidal behaviour following legislation to limit access to specific means of suicide (Box 2 and Figure 2). Routine national surveillance datasets (such as hospital episode data or health surveys) may be

sufficient for this purpose when the data are collected with sufficient frequency and consistency to permit an interrupted time series design,[56 57] the intervention is applied to a large population, and the 'signal' of the effect is large enough to be detected against the background 'noise' in the surveillance data.[58] For interventions at a more local level, routine data sources may not be suitable if they are collected infrequently (as in most national censuses), sample only a small fraction of the population, or involve exposure or outcome measures that are imprecise or change over time (as in some national health surveys)[59] or are made available at an insufficient level of spatial resolution (as in many datasets deposited in data archives). In the latter case, it may be possible to negotiate access to a version of the dataset in which more precise geocoding of participants is traded for less precise measures of their sociodemographic characteristics. If suitable routine data are not available, it will usually be preferable to collect baseline data and conduct a prospective (cohort or repeat cross-sectional) study rather than relying exclusively on retrospective (recall) data, although, as the example of sudden infant death syndrome (SIDS) shows,[54 60] retrospective case-control studies can contribute to important advances in public health research, interventions and policy.

4. The intervention or the principles behind it have the potential for replication, scalability or generalisability. This principle would most obviously be met when a proposal to roll out a given intervention across a whole region or country has political support and is informed by a well designed pilot study. However, it is not necessary for the precise intervention available for study to be 'replicable' as such in order to justify its evaluation, as long as the study is designed to address appropriate research questions. As the revised MRC framework for the development and evaluation of complex interventions makes clear, the evaluative focus on a given type of intervention may gradually shift over time from testing efficacy, to establishing practical effectiveness in a range of settings, to optimising delivery and understanding causal mechanisms.[61] For example, many interventions in the built environment are highly specific to their context. It is meaningless to suggest that a particular new railway, motorway, cycle path or country park could be directly replicated anywhere else, but it is possible to use the opportunity presented by natural experiments of this kind to test more general (and generalisable) hypotheses, particularly where multiple study sites or study populations are available.[62 63] Similarly, even if an overall programme of community-level interventions involves a high degree of internal complexity and variability between sites, it may be possible to identify elements within the overall programme that are similar across multiple sites and therefore offer more tractable hypotheses to be tested.

3.2. Natural and planned experiments

By definition, natural experimental studies involve a degree of opportunism: in ***The Uses of Epidemiology*** Jerry Morris called them 'experiments of opportunity'.[64] But making the most of the available opportunities requires careful planning. Many of the examples we have described use routinely collected data, such as mortality and health service records or population health surveys. Researchers need to be aware of the potential of such data sources, but also of their shortcomings, and should be ready to press for improvements where necessary. Haw et al.[17] describe a programme of studies to evaluate the impact of the Scottish ban on smoking in public places using a mixture of primary data collection, routine data and enhancements to existing sources, such as additional questions and measures incorporated into long established surveys (Box 4).

Box 4: Evaluating Scotland's smoke-free legislation

Legislation to ban smoking in enclosed public places was implemented in Scotland in March 2006. Evidence from previous research about the harm associated with environmental tobacco smoke (ETS), and the success of smoke-free legislation in the Republic of Ireland, were used to make the case in Scotland, and researchers worked closely with policy makers to ensure that the evidence was used to good effect. Compared with most public health interventions the legislation was a simple intervention, comprehensively implemented at a population level with an established infrastructure for enforcement. This made it an ideal policy to evaluate. However, the multiple outcomes across several domains required a complex evaluation strategy, again developed in collaboration with policy-makers. The strategy was based on an empirically derived logic model[17] linking the ban to short-term, intermediate and long-term outcomes. The evaluation focused on eight outcome areas: knowledge and attitudes; ETS exposure; compliance with the legislation; socio-cultural adaptation; smoking prevalence; tobacco-related morbidity and mortality; economic impacts on the hospitality sector; and health inequalities. Outcomes were measured using a combination of secondary analysis of routine datasets (health, behavioural and economic), social surveys and data from a portfolio of research studies commissioned to address specific questions.

The evaluation found that compliance was high from the outset[65] with a rapid[66] and sustained[67] improvement in the air quality in bars, with a 90% reduction in SHS smoke exposure in bar workers, as determined by salivary cotinine levels, and a

40% reduction in representative samples of adults[68] and children[69] in the general population. There were also measurable improvements in health. At one year the bar worker cohort reported fewer respiratory and sensory symptoms,[70] though no improvement in lung function. A prospective study of admissions to hospital for acute coronary syndrome found a 17% reduction in a ten month period post-legislation, compared with a 4% reduction in acute myocardial infarction (AMI) admissions in England over the same period and a mean annual reduction of 3% in AMIs in Scotland in the ten years prior to the legislation.[71] A time series analysis of routine admission data found a reduction in asthma hospitalizations in children of 18.2% per year. Prior to the legislation, admissions for asthma among children aged 0-14 had been increasing at a mean rate of 5.2% per year.[57]

In addition to the health impacts, the evaluation also found evidence of changes in smoking culture and behaviour. Post-implementation, support for the legislation increased more quickly in Scotland than in the rest of the UK where it had not yet been introduced. More stringent restrictions on smoking at home were also reported post-legislation in Scotland.[68 69 72] A time series analysis revealed that adult smoking prevalence fell by 2.4 percentage points after introduction of the legislation.[73] This was associated with an increase in uptake of nicotine replacement therapy in the three months leading up to implementation, though this fell back to the seasonal norm within a year of implementation.[73 74]

The research studies that contributed to the evaluation employed a variety of study designs. Evidence from this evaluation and those in other countries[75 76] makes a compelling case for the positive health, behavioural and socio-cultural impacts of the smoke-free legislation.

Researchers should also look for opportunities to incorporate an element of planned experimentation into the implementation of policies, especially in situations where a natural experiment is unlikely to provide convincing evidence. Murphy et al.[77] describe working with colleagues in the Welsh Assembly Government and with practitioners to ensure that the National Exercise Referral Scheme was rolled out in the context of an RCT, in line with the NICE recommendation that the evidence base was weak and such schemes should only be implemented in conjunction with further controlled studies.

Another option is to embed a trial within the implementation of a wider programme, to test the effectiveness of a variant or additional programme element. The Sure Start programme to tackle child poverty and social exclusion has been criticised for being rolled out nationally without an experimental evaluation.[78] Hutchings et al.[79] conducted a randomised controlled trial to determine the effectiveness of a parenting intervention delivered by Sure Start staff to families where there was a risk of a child developing a conduct disorder. In contrast to the ambiguous findings from the much larger evaluation of Sure Start,[25 80] the trial showed that the parenting intervention was effective[79] and good value for money.[81]

Stepped wedge designs, in which the intervention is rolled out area by area with the order decided randomly, offer a compromise between non-experimental implementation and a conventional cluster randomised trial.[82] They provide another way of incorporating an experimental element within programme implementation and are particularly useful in situations, such as Sure Start, where implementation is likely to be phased in any case, or where existing, partial evidence of effectiveness makes a parallel group randomised design politically or ethically unappealing.

3.3. Assessing value for money

Given limited budgets, policy-makers face difficult choices regarding which interventions to fund, and to what extent. Evidence of effectiveness, while necessary, is not sufficient to make properly informed choices about the best use of scarce resources. Economic evaluation uses information about both costs and outcomes to estimate value for money as accurately as possible given the available evidence. Incorporating an economic perspective into a natural experimental evaluation should make the results much more useful for decision-makers, helping them to choose between interventions offering the greatest value for money — that is, to select the best intervention to meet a particular objective ("technical efficiency") and to allocate funds between different kinds of interventions ("allocative efficiency").

Ideally, economic evaluation should be considered as an integral part of the overall evaluation,[83] although in some cases it may be more appropriate to apply economic evaluation later, when synthesising and interpreting the results of multiple intervention studies, rather than separately for each study. Economists should be involved at project inception to ensure the right information is collected and a value of information analysis carried out at an early stage to assess how much decision-makers should be prepared to pay for the evidence, and thus inform the decision whether to proceed with the evaluation.[84]

In practice, methods of economic evaluation have largely been developed in relation to healthcare rather than public health interventions, and it is recognised that economic evaluation in public health poses particular challenges.[85] These include the attribution of effects, the comparison of interventions originating in different policy sectors, the measurement of outcomes and the incorporation of equity considerations.[86] The first three are likely to be particularly pronounced in the case of natural experiments. Section 4 reviews

methods for estimating effects; here we briefly discuss the questions of which perspective to adopt and the choice of cost and outcome measures.

Whose perspective? A key feature of economic evaluation is that it makes explicit the perspective from which cost and impact are assessed. The perspective adopted then informs which costs and outcomes to include in the evaluation. Choice of perspective is complicated in cases where health impacts are not the primary aim of the intervention. In the case of natural experiments, a broad, eg societal, perspective will usually be preferable to a narrower one, such as the perspective of a service provider.[87]

Which costs and which outcomes? A wide range of costs may have to be considered, including the costs of providing the intervention, the impact on the use of other services, and costs associated with unintended consequences, such as the economic impact of licensing laws, minimum prices per unit of alcohol, or bans on smoking in bars or restaurants. Costs should be calculated net of any savings (eg reduced hospitalisations) that occur as a consequence of the intervention. Use of an explicit theory of change or a logic model may help to systematically identify all the relevant costs and outcomes, and is particularly useful in identifying long term consequences. Estimating long term impacts is likely to be heavily reliant on untestable, or only partly testable, assumptions and should therefore be accompanied by a sensitivity analysis. Generic, preference weighted, outcome measures should be used where appropriate, reflecting the perspective taken in the evaluation. This enables heterogeneous interventions to be compared, and value for money decisions to be taken.[85]

4. Improving the design, analysis and reporting of natural experiments

Planned and natural experiments are both subject to similar threats to validity, such as loss to follow-up, inaccurate assessment of exposure and outcomes, and so on. The key difference is that randomised controlled trials have a very general and (if deployed properly) effective method of preventing the bias that results from selective exposure to the intervention,* i.e. the tendency for exposure to vary according to characteristics of participants that are also associated with outcomes.[88] In the case of non-randomised studies, there is no such general solution to the pervasive problem of confounding.[89] Instead there is a range of partial solutions, each if which is applicable in some, often very restricted, circumstances but not others. Understanding the processes that produce the variation in exposure is therefore critical to the design of natural experimental studies.[90]

All natural experimental studies require a comparative design of some kind, to provide an indication of what would have happened in the absence of the intervention, often referred to as a 'counterfactual'. If an intervention is expected to have a very large or rapid impact, a simple design, such as a comparison of outcomes in an exposed group and an unexposed control group, may appear to be adequate. Natural experiments can also be used to study more subtle effects, so long as a suitable source of variation in exposure can be found, but the design and analysis become more challenging. In any case, what is often required from an evaluation is an estimate of effect size, and a large observed effect may incorporate a large element of bias due to selective exposure to the intervention. Whatever the expected effect size, care should be taken to minimise bias in the design and analysis of natural experiments. Since it is difficult to eliminate all bias, transparent reporting of natural experimental studies is vital.

4.1. Design

A study protocol should be developed whatever design is adopted. Interventions evaluated using natural as opposed to planned experiments may be vaguely specified, prone to delays in formulation or implementation, and subject to unplanned variation or change over time.[91] A credible study protocol, developed in collaboration with the relevant policy makers or service providers, can help avoid some of these difficulties. Publication of study protocols helps to publicise research in progress, thus reducing the risk of unhelpful duplication and selective publication of positive results.[92] It makes clear what the initial study hypotheses were, which analyses were pre-planned, and which, if any, were adopted following inspection of the data. Other aspects of good practice in the conduct of observational studies that are particularly important in the case of natural experiments include the need for clear definitions of target populations, explicit sampling criteria, and valid and reliable measures of exposures and outcomes.

The examples of suicide prevention[18 38] and air pollution control[42 43] show that simple designs can provide convincing evidence if a whole population of considerable size is abruptly exposed to an intervention, the effects of the intervention are large and rapidly follow exposure, and can be measured accurately at population level using routinely available data. This combination of circumstances is rare, and more complex designs are usually required.

Multiple pre/post measures: If it is not feasible to include an unexposed control group, as in the alcohol taxation example (Box 3), then repeated measures before and after the intervention may be used to control for secular changes, as in an interrupted time series design.[56] These should preferably extend long enough before and after the intervention to taken into account any short term fluctuations around the time of implementation.

Multiple exposed/unexposed groups: Intervention and control groups should ideally be as similar to one another as possible. If there are substantial differences, for example due to strong selection into the intervention group, complete control for confounding will be difficult. Multiple comparison groups that differ according to some variable that may affect both exposure to the intervention and outcomes can be used to assess whether selection on that variable is likely to be an important source of bias.[90] Variations in policy, or in policy implementation, between states or provinces in a federal jurisdiction may provide a useful source of variation, especially if the areas being compared are similar in the terms of other characteristics that might affect response to the policy.

* This is sometimes referred to as selection bias or allocation bias. The term 'selection bias' is used in two distinct senses within epidemiology.[15 84] One is the sense used above, i.e. to refer to the likelihood that exposure varies with factors that are also associated with variation in outcomes; the other refers to a bias arising from the process by which a study sample is selected from the relevant population. Selection bias in the first sense causes confounding, which can reduce 'internal validity' and is a key problem in interpreting natural experimental data. Selection in the other sense is also important, because it affects the generalisability or 'external validity', of study findings, but is less pivotal to the methodology of natural experimental studies.

Measurement of confounders: the methods of analysis and testing described in the next section rely heavily on accurate identification and measurement of potential confounders, i.e. characteristics of participants that are associated with both exposure and outcomes. Ideally, the choice of measures should be based on a good theoretical understanding of the selection processes that are causing the confounding. Often, however, the choice depends on what is available in routine or pre-existing survey datasets.

Combinations of methods: It is unlikely that a single method will deal adequately with all possible biases, and combining methods may provide additional protection. Controlled interrupted time series designs combine multiple pre-post measures with comparisons of exposed and unexposed groups. Cohort studies are more powerful than repeat cross-sectional studies, for a given sample size, but also more subject to biases due to ageing, attrition, etc., so combining the two designs may be preferable to using either in isolation.[93] Mixed method approaches, combining qualitative and quantitative methods may also be useful, for example where it is difficult to distinguish quantitatively between competing explanations.

4.2. Analysis

One of the defining features of a natural experiment is that manipulating exposure to the intervention is not possible. Understanding the assignment process is therefore central to the appropriate analysis and interpretation of data from natural experiments.[90] There are a few examples where assignment is by a 'real life' lottery, but selection is the rule and a range of methods is available for dealing with the resulting bias. This is an active area of methodological research, and many variants and extensions are described in the specialist literature.[12 19 94 95] Here we summarise some of the most commonly used methods. It is important to emphasise that, as far as possible, analytical options need to be taken into account in the design of studies to ensure that data requirements are met.

4.2.1. Selection on observables

One important class of methods is applicable where the factors that determine exposure can be measured accurately and comprehensively.

Matching: This involves finding unexposed individuals (or clusters of individuals) which are similar to those receiving the intervention, and comparing outcomes in the two groups (Box 5).

Box 5: A large scale natural experiment: incentives for Indian women to use a health facility to give birth

Trials of conditional cash transfer schemes in low and middle income countries suggest that they can increase uptake of preventive services, but their effect on health is less clear.[96] Janani Suraksha Yojana (JSY) is a conditional cash transfer scheme launched by the Government of India in 2005 to encourage women to give birth in a health facility rather than at home, with the aim of reducing maternal and neonatal deaths.

Lim et al.[97] used data from nationwide household surveys in 2002-4 and 2007-8 to assess the coverage of JSY and its effect on use of health facilities and rates of perinatal, neonatal and maternal mortality. Three methods were used to assess impact: exact matching, in which births to mothers receiving JSY were matched with non-recipient births, using measures of poverty, wealth, caste, education, parity, maternal age and place of residence, with further adjustment in the analysis for those covariates plus others including religion and distance from the nearest health facility; a 'with vs. without' analysis, comparing all births to women receiving JSY with all those to non-recipients, adjusted for the same socioeconomic and demographic characteristics; and an analysis of 'difference in differences' that compared use of health facilities and mortality at a district level, adjusting for a range of social, demographic and health system variables.

Analysis of coverage showed wide district and state-level variation in the take-up of JSY. States with higher rates of take-up showed larger increases in the proportion of births within health facilities. All three analyses of impact showed that women who received payments were more likely to give birth in a health facility. The matched and 'with vs. without' analyses indicated that JSY receipt was associated with four fewer perinatal deaths per 1000 pregnancies and two fewer neonatal deaths per 1000 live births. The less precise difference-in-differences analysis showed a larger but non-statistically significant reduction in perinatal and neonatal deaths at the district level.

Strengths of this study include the combination of three methods to assess impact, using data from very large surveys (620,000 households in the first wave and 720,000 in the second), which allowed analyses of state and district

level variation. The matched and 'with vs. without' analyses are prone to confounding by unobserved individual level differences. The difference-in-differences analyses address this problem, but are less precise. The consistent results in relation to intervention coverage, and the consistent direction of effect in the neonatal and perinatal mortality analyses, suggest that differences in the mortality findings reflect limited statistical power in the district-level analyses rather than bias in the individual-level analyses. As with other successful natural experimental studies, the study demonstrates the value of high quality data gathered across large populations.

Matching requires good quality data on the relevant characteristics and a sizeable unexposed population of potential matches. An advantage is that it does not require data on changes in outcomes. A drawback is that the matching process may become unwieldy if there are a large number of factors to take into account, though specialised software is now available.[98] However the main disadvantage is that interpreting the difference in outcomes as an effect of the intervention assumes that selection takes place only on observable characteristics. Bias will remain if there are unobserved factors that influence both exposure and outcomes.

Regression adjustment: Measured characteristics that differ between those receiving the intervention and others can be taken into account in multiple regression analyses. Issues that arise include the choice of which variables should be adjusted for and whether this should be pre-specified, and whether to include non-linear terms and interactions in the statistical model. Such regression adjustment is only effective if all the factors that determine exposure are precisely measured. This may very often not be the case,[89] in which case estimates of an intervention's impact will be biased by residual confounding.

Propensity scores: In formal terms, a propensity score is an estimate of the likelihood of being exposed given a set of covariates.[99] The scores are usually estimated by logistic regression, and can be used in a variety of ways. One is to match exposed with unexposed units (which may be individuals or clusters of some kind) using values of the propensity score rather than the covariates themselves. Another is to compare exposed and unexposed units within strata of the propensity score. They can also be used for covariate adjustment, by entering the score into a regression of outcomes. The advantage is that using a single score rather than a range of covariates should make it easier to find matches, or to keep the number of strata or the number of variables in the outcome model manageable. A disadvantage is that propensity scores only work well when there is substantial overlap between the scores of exposed and unexposed units. Like matching, the method cannot address the bias associated with unobserved differences between recipients and non-recipients.

4.2.2. Selection on unobservables

Given the difficulty of measuring accurately all of the characteristics associated with exposure to an intervention, methods that deal with unobserved factors are a potentially valuable advance on those that only deal with observed factors.

Difference in differences: This method compares change over time in exposed and unexposed groups.[12] The differencing procedure controls for unobserved individual differences, and for common trends, i.e. changes that affect both groups similarly (Box 6). Because it assumes that the unobserved characteristics are fixed, and that the outcomes in each group would change in the same way in the absence of the intervention, it is vulnerable to changes in the composition of the groups and to external influences, such as the effect of other interventions, that differentially affect the exposed and unexposed groups. With additional data it may be possible to address these problems. Compositional changes can be allowed for with individual level data, and the assumption that trends would be similar in the absence of the intervention can in principle be tested using data on multiple time periods.

Box 6: Abolition of general practitioner fundholding as a natural experiment

There was no large-scale evaluation when general practitioner (GP) fundholding was introduced as part of a series of changes in National Health Service funding in the early 1990s, despite concerns that it might lead GPs to accumulate surpluses by referring fewer patients to hospital for elective treatment. A feature of the scheme was that GPs could choose whether or not to become fundholders, and around half did so, meaning that comparisons of practices inside

and outside the scheme might be confounded by factors associated with their choices. The complete withdrawal of fundholding in 1999 provided the opportunity for a natural experiment, comparing changes in admission rates in fundholders and non-fundholders before and after abolition.

Dusheiko et al.[100] used a difference-in-differences approach to identify the effect on admission rates of fundholding, allowing for differing characteristics of fundholding and non-fundholding practices and for other influences on admissions, such as initiatives to reduce waiting times. Using a mixture of administrative and research datasets with information on referral rates and practice characteristics, they found that fundholders had lower rates of elective admissions while fundholding was in operation, and subsequently increased their rates of admission more than non-fundholders in the two years following abolition. There was no difference in rates of emergency admission between the two types of practice, or in the changes in the rates before and after abolition, strengthening the inference that fundholding influenced elective admissions.

The study has a number of features that make for a good natural experimental study. It made use of large national datasets, with information on admissions, practice and provider characteristics, that permitted modelling of factors other than fundholding that might affect admission rates; it used data for two years before and after the withdrawal of fundholding, to distinguish longer term from transitional effects; there was an abrupt change in financial regime that affected all the fundholding practices, but had no effect on the non-fundholders; it used a difference-in-differences approach to deal with confounding due to factors associated with participation decisions, and a non-equivalent dependent variable (emergency admissions) to test the assumption that unobserved temporal factors affected the two groups of practices similarly.

Instrumental variables: An instrumental variable (IV) is a factor associated with exposure to an intervention, but independent of other factors associated with exposure, and associated with outcomes only via its association with exposure (an assumption known as the 'exclusion restriction'). In a well-designed RCT, treatment allocation satisfies these requirements because it is random and therefore independent of the characteristics of participants, and is associated with outcomes only via its association with receipt or non-receipt of treatment. A good example in a non-experimental context is 'Mendelian randomisation'[101] - the use of a genetic variant (for example in a gene that controls alcohol metabolism) that mimics the effect of an environmental exposure (alcohol consumption), but is uncorrelated with other characteristics that may go together with the exposure of interest (such as smoking). IVs have also been used in 'outcomes research' to evaluate the impact of treatment using routine data.[102] In these studies, variables such as distance from specialised centres have been used to evaluate novel treatments, the assumption being that patients living close to a specialised centre are more likely to receive the novel treatment, but are otherwise similar to other patients.[103] Stukel et al.[104] used regional variations in catheterisation rates as an instrument to estimate the effect of cardiac catheterisation on patient survival, reasoning that the regional rates were unlikely to be associated with individual prognostic factors. They compared the results from the IV model with those of survival effects estimated using conventional multivariable and propensity score-based risk adjustment. The IV estimates were markedly lower than those of the other models, and closer to the effect sizes found in randomised trials.

IVs are a potentially powerful method for dealing with confounding, but good instruments are scarce.[95 102] A weak instrument, i.e. one that is only weakly associated with exposure, will tend to overestimate the impact of the intervention, and violation of the key assumptions will also lead to bias. The validity of the assumptions can be explored, but is hard to prove, and it has been suggested that IVs replace the unverifiable assumption in conventional methods that there is no unmeasured confounding, with equally untestable assumptions about the properties of the instrument.[105] Instruments should therefore be chosen on the basis of a good theoretical understanding of their relationship with other variables associated with exposure and outcomes.

Regression discontinuity designs: In its basic form this approach exploits a step change or 'cutoff' in a continuous variable used to assign treatment, or otherwise determine exposure to an intervention. The assumption is that units (individuals, areas, etc.) just below and just above this threshold will otherwise be similar in terms of characteristics that may influence outcomes, so that an estimate of treatment effect can be obtained by comparing regression slopes either side of the cutoff. Regression discontinuity designs have been widely used in economics and education research, and there is an extensive literature that deals with the methodological issues.[12 19 106] A limitation of the approach is that it only approximates to a randomised comparison for units close to the threshold. Another is that bias may be reintroduced if individuals react to their assignment. As with IV approaches, its applicability is somewhat limited but the targeting of programmes by income, poverty and other continuous traits provides opportunities (Box 7).

Box 7:
Did Headstart reduce mortality? – example of a regression discontinuity design

The US Headstart programme has been extensively researched but uncertainties remain about the size and durability of its effects. When the programme was first implemented, help with applying for funding was targeted on the 300 poorest counties, to prevent them from losing out in the competition for funds. As a result, 80% of the poorest counties received Headstart funding, compared with 40% of all counties, and their level of funding per child was twice the level in counties with poverty rates slightly below the cutoff used to assign help with funding applications. Ludwig and Miller[107] use this source of variation as the basis for a regression discontinuity approach to evaluating the impact of Headstart on child health and educational outcomes.

Headstart provided parenting support, nutrition, social and health services – including screening and immunisation programmes - as well as help with schooling. To identify its health impact, Ludwig and Miller compared regressions of mortality on poverty rates for the 228 counties with rates 0-10% above the cutoff, and 349 counties with rates of 0-10% below the cutoff, for a set of Headstart-related causes of death. For those causes of death they identified a discontinuity in mortality rates at the cut off, equivalent to 1-2 fewer deaths per 100,000 children.

They tested this finding in a number of ways. First, the effect was restricted to Headstart-related causes of death, and was only seen in children whose ages made them eligible for Headstart services. It was unlikely to reflect other help for the poorest counties, as there was no discontinuity in non Headstart-related funding at the cutoff. Selective migration is another possible explanation, but migration rates were low and there was no discontinuity in population characteristics to suggest that selective migration was producing spurious gains. This rigorous approach to testing, as well as the transparent, well-documented source of variation in access to programme funding makes this a good natural experiment.

Neidell[108] used an interesting variant of this approach to estimate the extent to which people respond to smog warnings triggered by air pollution. As pollution is associated with other potentially unpleasant features of the weather, such as high temperatures and humidity, changes in behaviour associated with smog warnings may reflect reactions to the weather, rather than to the warnings. Neidell compared participation in outdoor activities on days when air pollution levels were just below or above the levels that triggered different types of warning. While participation changed at the threshold, there was no discontinuity in relevant covariates, suggesting that people responded to the warnings rather than to the weather.

4.3. Strengthening causal inference

In practice, none of these approaches provides a comprehensive solution to the central problem of selective exposure to the intervention.[89] Methods of controlling for observed factors associated with receipt of treatment are vulnerable to selection on unobservables. Methods for dealing with selection on unobservables require strong but untestable assumptions and are restricted in their application by the availability of good instruments. These methods are therefore best used in conjunction with additional tests for the plausibility of any causal inferences.

Information on mediators of change: Information on links in the causal chain between intervention and outcome can strengthen confidence in attributing changes to the intervention. In their evaluation of the impact of the Scottish ban on smoking in public places on hospitalisations for acute coronary syndrome, Pell et al.[71] (Box 4) collected information on exposure to secondhand smoke as well as active smoking – a marked improvement on most other studies of similar interventions.[75] Likewise Hedley et al[43] and Clancy et al[42] were able to show that levels of the relevant pollutants fell following the bans on high sulphur fuels and domestic coal sales respectively.

Non-equivalent dependent variables: Changes in outcomes that are not expected to respond to the intervention can be used to assess specificity of effect. If related and unrelated outcomes change in a similar way, it is less plausible to attribute change to the intervention, as opposed to secular trends or the effect of some other intervention. Dusheiko et al[100] (Box 6) used emergency admissions to test whether changes in elective admissions could plausibly be attributed to GP fundholding. Ludwig and Miller[107] (Box 7) compared mortality from causes that might respond to immunization and screening, with mortality from causes that were unlikely to be affected.

Combining methods and comparing results: Combining analytical methods, especially those that address differing sources of bias, is a potentially powerful way of exploring the dependence of results on the key assumptions of the different methods, strengthening causal inference and generating more robust estimates of effect size and direction.[104] In the JSY evaluation (Box 5), Lim et al.[97] combined methods for dealing with selection on observable and unobservable factors that might be associated with participation. Likewise, Stukel et al compared estimates from IV models, propensity-score based models and conventional multivariable adjustment.[104] Belot and James combined a difference in difference analysis with propensity score matching based to analyse the effect of participation in a television campaign to promote healthy school meals on educational outcomes. Schools taking part were matched with those in neighbouring education authorities. Absenteeism fell and educational attainment improved in the intervention schools.[109]

Sensitivity analysis: Complete control for confounding is unlikely in a non-randomised study, and the success of efforts to deal with confounding is hard to determine. Sensitivity analyses can be used to assess the potential importance of unmeasured confounders, and other potential sources of bias such as loss to follow-up, missing data on exposure and outcomes, etc. A number of methods have been developed that can give an indication of the effect of an unmeasured confounder given its prevalence and strength of association with exposure and outcome.[110]

Replication: Given the difficulty of eliminating bias, single studies are unlikely to be definitive, and replication is needed to build up confidence in conclusions about effectiveness. Exact replication of a natural experiment is unlikely, but partial replication is often possible and may be more informative. Consistent findings from studies using varying designs makes it less likely that common biases are present, and consistent findings across settings or populations increase confidence in the generalisability of causal inferences. A number of studies in different countries have shown that legal restrictions on smoking in public places reduce hospital admissions for heart attacks. Although the size of the effect varies widely, as might be expected given variation in the prevalence of active smoking and the extent of partial restrictions prior to outright bans, the predominantly positive results suggest a real effect.[75 76]

4.4. Reporting natural experiments

As mentioned above (Section 4.1) protocols for natural experimental studies should be published. In relation to results, guidelines have been developed for reporting observational studies, along similar lines to the influential CONSORT guidelines for reporting randomised trials. The STROBE checklist provides a good basis for reporting natural experiments. It is referred to in the Uniform Requirements for Manuscripts Submitted to Biomedical Journals by the International Committee of Medical Journal Editors, and is periodically updated (www.strobe-statement.org).[111] The TREND statement,[112] which is specifically designed for reporting non-randomised intervention studies, is largely consistent with STROBE but has been less widely adopted.

Items that require particular attention in the reporting of natural experiments include:

Title/abstract – reports of natural experimental studies often do not identify the study as one that uses a natural experiment. We recommend that the term is used in the abstract to aid future searching, alongside terms that identify population, intervention, comparator, outcome, etc. For the main report of study methodology, it should also be used in the title.

Background/rationale – it is crucially important to provide a clear description of the intervention, with links and/or references to a more detailed description, plus a summary of any interventions received by any control or comparison groups. It is also useful to explain why a natural experimental approach has been followed.

Study design – again, to aid future searching it is useful to explicitly state the method used to estimate impact, using standard terminology. The assignment process, i.e. the mechanism that determines exposure to the intervention, should be clearly described and the unit of assignment (individual, school, GP practice, town, etc.) indicated.

Outcomes – wherever possible, effects should be presented in natural units, eg reduction in numbers of events, or changes in absolute risks, rather than just in terms of relative risk reductions, etc.

Limitations – the extent to which the assignment process is selective in ways likely to cause bias, the effectiveness of any post-hoc adjustment, and the strength and direction of any remaining bias, should be discussed in a detailed and balanced way. Ideally, qualitative judgements about the risk of bias, and how well it has been dealt with, should be supplemented by a

quantitative assessment of the impact of loss to follow-up, exposure misclassification, unmeasured confounders, missing data, etc., derived from sensitivity analyses.[113 114]

Interpretation – if the study used multiple methods, variation in the estimates from the different methods should be highlighted and, as far as possible, explained. If possible, the results should be compared with those of other evaluations, paying attention to any associations between features of the intervention, evaluation methods and effect sizes.

4.5. Systematic reviews of natural experiments

Systematically reviewing the evidence from natural experimental studies is demanding, because of the variety of designs, the difficulty of setting sensitive and specific search criteria and the wide array of potential sources of bias.[115] The registration, conduct and reporting of clinical trials is increasingly standardised and regulated, but the process has barely begun in relation to non-randomised studies. Despite the problems, systematically reviewing the evidence from natural experimental studies is important to

- Inform research priority setting and identify promising interventions for further development and evaluation
- Aid the interpretation of new evidence, for example in the discussion section of papers reporting natural experiments
- To provide 'best available' estimates of intervention effectiveness in areas where only observational evidence is available or where natural experiments predominate, such as the effects of price changes on alcohol consumption[45 46] or the effect of secondhand smoke exposure on cardiovascular disease.[75]

It follows that there is a range of audiences for the synthesised evidence, whose differing requirements need to be taken into account in deciding on the best approach.[116] The Cochrane Non-Randomised Studies Methods Group has developed guidance for including non-randomised studies in systematic reviews.[117] It recommends that where methods, interventions or settings vary widely, studies should be grouped into a series of component reviews rather than a single review. On the other hand, the Cochrane Public Health and Health Promotion Field notes that a single review addressing a broader question will better inform decisions about which interventions to implement from among a range of options and may therefore be of more use to policy makers.[118] Meta-analysis should only be used where there are a reasonable number of high quality studies using a similar design. It is inadvisable for studies using different designs or design elements, or where study quality varies widely. While this approach will yield the most precise, unbiased estimates of effect, some users of the evidence may prefer a less stringent approach that keeps more of the available evidence in play.

There are a wide range of tools available for assessing study quality, and the Cochrane Group recommends the 8-item Newcastle-Ottawa scale,[119] partly on grounds of ease of use, which is an important consideration if a large number of studies need to be screened. Risk of bias should be assessed according to design features, such as whether a sensitivity analysis was carried out, etc., rather than on the basis of broad labels, such as cohort study, cross sectional study, etc. Particular attention should be paid to the methods used to adjust for confounding. Graphical methods, such as forest or funnel plots, are preferable to narrative synthesis alone, because they may reveal study heterogeneity or publication bias. Studies relying on secondary analysis of existing datasets may emerge gradually from an ongoing programme of work, only becoming discrete studies when an 'interesting' association is found, thus heightening the risk of publication bias. Finally, results should be interpreted cautiously: large effects do not mean that bias is unimportant, as there may be a large bias component in the estimate.

5. Conclusion

Natural experimental approaches have much to offer, but it would be unwise to assume they are suitable for evaluating interventions when planned experiments are impractical, without very detailed consideration of the kind of opportunities the intervention will generate. Over-optimism about natural experiments should not be allowed to deflect attention from the need and opportunity to conduct RCTs of population health interventions. Often, an RCT will be the only way to obtain reliable estimates of the impact of an intervention, especially if exposure is likely to be highly selective in the absence of randomisation. Even so, natural experimental studies need not be restricted to situations where the expected effects are large. They can be used to detect more subtle effects where there is a transparent exogenous source of variation provided by the sudden introduction or withdrawal of an intervention in a whole population, or by some other feature of the process that determines exposure. Understanding this process is key to designing appropriate natural experimental studies.

There are important areas of public health policy where natural experiments have already contributed a convincing body of evidence. What is needed in order to make best use of natural experiments in future? We believe that the following six issues are particularly important.

First, good working relationships between researchers and policy makers and flexible forms of research funding are necessary in order to exploit the opportunities generated by policy changes. We recognise that there may be increased costs and risks involved, but these may be justified where the interventions themselves are costly or strategically important. In situations where the obstacle to experimental evaluation is political rather than scientific, researchers should work with policymakers to identify ways in which random allocation or other design elements might be introduced into new programmes to permit the use of more robust intervention study designs.

Second, research effort should focus on important but answerable questions, accepting that some interesting questions may be genuinely intractable, and taking a pragmatic approach based on combinations of methods, careful testing of assumptions and transparent reporting.

Third, given the difficulty of eliminating bias in non-randomised studies, quantitative estimates of bias should become a standard feature of reporting.

Fourth, a prospective register of natural experimental studies, as has already been suggested in the case of smoke-free legislation[76] and for public health interventions generally,[92] would be a major step forward.

Fifth, the case studies we have presented illustrate the crucial role of routinely-collected data, either via administrative systems or long-running population surveys. If we are to rely more on natural experiments, investment in improving linkage is needed, both between health data sets and across heath, education, social security and other data sources. Researchers should take opportunities to argue for the enhancement and linkage of national survey datasets to optimise their utility for studies of future natural experiments.

Finally, public health can learn from other disciplines faced with similar evaluation challenges. There are a number of promising methods that have been little used to evaluate population health interventions to date. Building up experience of these promising but lesser used methods, to determine whether and in what circumstances their theoretical advantages and disadvantages matter in practice, is another key to future progress.

Bibliography

1. Wanless D. ***Securing good health for the whole population***. London: HM Treasury, 2004.
2. Foresight. ***Tackling Obesities: Future Choices***. London: Department for Innovation, Universities and Skills, 2007.
3. Department of Health. ***Healthy Weight, Healthy Lives: a research and surveillance plan for England***. London: Department of Health, 2008.
4. Robinson G, McNulty JE, Krasno JS. Observing the counterfactual: the search for political experiments in nature. ***Political Analysis*** 2009;17:341-57.
5. Rosenzweig MR, Wolpin KI. Natural 'natural experiments' in economics. ***Journal of Economic Literature*** 2000;38(4):827-874.
6. Diamond J, Robinson JA, editors. ***Natural Experiments of History***. Cambridge, Mass: The Belknap Press, 2010.
7. Dunning T. Improving causal inference: strengths and limitations of natural experiments. ***Political Research Quarterly*** 2008;61:282-92.
8. Angrist J, Lavy V. Using Maimonides' rule to estimate the effect of class size on scholastic achievement. ***Quarterly Journal of Economics*** 1999;114:533-75.
9. Diamond J. Intra-island and inter-island comparisons. In: Diamond J, Robinson JA, editors. ***Natural Experiments of History***. Cambridge, Mass: The Belknap Press, 2010.
10. Susser E, Hoek HW, Brown A. Neurodevelopmental disorders after prenatal famine. The story of the Dutch Famine Study. ***American Journal of Epidemiology*** 1998;147(3):213-6.
11. Sekikawa A, Horiuchi BY, Edmundowicz D, Ueshima H, Curb JD, Sutton-Tyrrell K, et al. A natural experiment in cardiovascular disease epidemiology in the early 21st century. ***Heart*** 2003.
12. Angrist J, Pischke J-S. Mostly harmless econometrics: an empiricist's companion. Princeton: Princeton University Press, 2009.
13. Davey Smith G. Commentary: Behind the Broad Street pump: aetiology, epidemiology and prevention of cholera in mid 19th century Britain. ***International Journal of Epidemiology*** 2002;31:920-32.
14. Diamond J, Robinson JA. Natural experiments: working in the history lab. ***New Scientist*** 2010(2753).
15. Last J, editor. ***A dictionary of epidemiology. 3rd ed***. New York: Oxford University Press, 1995.
16. Susser M. The logic in ecological: II. The logic of design. ***American Journal of Public Health*** 1994;84:830-35.
17. Haw SJ, Gruer L, Amos A, Currie CE, Fischbacher C, Fong GT, et al. Legislation on smoking in enclosed public places in Scotland: how will we evaluate the impact? . ***Journal of Public Health*** 2006;28(1):24-30.
18. Gunnell D, Fernando R, Hewagama M, Priyangika W, Konradsen F, Eddleston M. The impact of pesticide regulations on suicide in Sri Lanka. ***International Journal of Epidemiology*** 2007;36:1235-1242.
19. Shadish WR, Cook TD, Campbell DT. ***Experimental and quasi-experimental designs for generalised causal inference***. Boston, Mass: Houghton Mifflin Company, 2002.
20. Kelly MP, Morhan A, Bonnefoy J, Butt J, Bergman V, for the measurement and Evidence Knowledge Network. The social determinants of health: developing an evidence base for political action. Final report to World Health Organisation Commission on the Social Determinants of Health: National Institute for Health and Clinical Excellence, UK/Universidad de Desarollo, Chile, 2007.
21. West SG, Duan N, Pequegnat W, Glaist P, Des Jarlais DC, Holgrave D, et al. Alternatives to the randomised controlled trial. ***American Journal of Public Health*** 2008;98(8):1359-66.
22. Ogilvie D, Mitchell R, Mutrie N, Petticrew M, Platt S. Evaluating health effects of transport interventions: methodologic case study. ***American Journal of Preventive Medicine*** 2006;31(2):118-26.
23. Bonell CP, Hargreaves J, Cousens S, Ross D, Hayes R, Petticrew M, et al. Alternatives to randomisation in the evaluation of public health interventions. ***Journal of Epidemiology and Community Health*** 2011;65:582-87.
24. House of Commons Health Committee. Health Inequalities. Third Report of Session 2008-9. Volume 1 Report, together with formal minutes. London: The Stationery Office, 2009.
25. Melhuish E, Belsky J, Leyland AH, Barnes J, National Evaluation of Sure Start Research Team. Effects of fully established Sure Start Local Programmes on 3 year old children and their families living in England: a quasi-experimental observational study. ***The Lancet*** 2008;372:1641-7.
26. Academy of Medical Sciences. Identifying the environmental causes of disease: how should we decide what to believe and when to take action. London: Academy of Medical Sciences, 2007.
27. Rutter M. Proceeding from observed correlation to causal inference. The use of natural experiments. ***Perspectives on Psychological Science*** 2007;2(4):377-95.
28. Skivington K, McCartney G, Thomson H, Bond L. Challenges in evaluating welfare to work interventions: would an RCT design have been the answer to all our problems? ***BMC Public Health*** 2010;10:254.
29. Oakley A. Experimentation and social interventions: a forgotten but important history. ***British Medical Journal*** 1998;317:1239-42.
30. Rutter M. Environmentally mediated risks for psychopathology: research strategies and findings. ***Journal of the American Academy of Child and Adolescent Psychiatry*** 2005;44(1):3-18.
31. Costello EJ, Compton S, Keeler G, Angold A. Relationships between poverty and pyschopathology: a natural experiment. ***Jounal of the American Medical Association*** 2003;290(15):2023-29.

32. Costello EJ, Erkanii A, Copeland W, Angol A. Association of family income supplements in adolescence with development of psychiatric and substance use disorders in adulthood among an American Indian population. ***Jounal of the American Medical Association*** 2010;303(19):1954-60.
33. Geddes J. Suicide and homicide by people with mental illness. ***British Medical Journal*** 1999;318:1225-6.
34. Peto R, Baigent C. Trials: the next 50 years. ***British Medical Journal*** 1998;317:1170-1.
35. Taylor S, Kingdom D, Jenkins R. How are nations trying to prevent suicide? An analysis of national suicide prevention strategies. ***Acta Psychiatrica Scandinavica*** 1997;95(6):457-63.
36. Hawton K, Bergen H, Simkin S, Brock A, Griffiths C, Romeri E, et al. Effect of withdrawal of co-proxamol on prescribing and deaths from drug poisoning in England and Wales: time series analysis. ***British Medical Journal*** 2009;338:b2270.
37. Wheeler B, Gunnell D, Metcalfe C, Martin R. The population impact on incidence of suicide and non-fatal self harm of regulatory action against the use of selective serotonin reuptake inhibitors in under 18s in the United Kingdom: ecological study. ***British Medical Journal*** 2008;336:542-5.
38. Kreitman N. The coal gas story. United Kingdom suicide rates, 1960-71. ***British Journal of Preventive Medicine*** 1976;30:86-93.
39. Wheeler B, Metcalfe C, Martin R, Gunnell D. International impacts of regulatory action to limit antidepressant prescribing on rates of suicide in young people. ***Pharmacoepidemiology and Drug Safety*** 2009;18:579-88.
40. Chang S, Gunnell D, Sterne J, Lu T, Cheng A. Was the 7 economic crisis 1997-1998 responsible for rising suicide rates in East/ Southeast Asia? A time-trend analysis for Japan, Hong Kong, South Korea, Taiwan, Singapore and Thailand. ***Social Science and Medicine*** 2009;68:1322-31.
41. Stuckler D, Basu S, Suhrcke M, Coutts A, McKee M. The public health effect of economic crises and alternative policy responses in Europe: an empirical analysis. ***The Lancet*** 2009;374:315-23.
42. Clancy L, Goodman P, Sinclair H, Dockery DW. Effect of air pollution control on death rates in Dublin, Ireland: an intervention study. ***Lancet*** 2002;360:1210-14.
43. Hedley A, Wong C, Thach T, Ma S, Lam T, Anderson H. Cardiorespiratory and all-cause mortality after restrictions on sulphur content of fuel in Hong Kong: an intervention study. ***Lancet*** 2002;360(9346):1646-52.
44. Wells N. ***Suicide and Deliberate Self-Harm***. London: Office of Health Economics, 1981.
45. Booth A, Meier P, Stockwell T, Sutton A, Wilkinson A, Wong R. Independent review of the effects of alcohol pricing and promotion. Part A Systematic reviews. Sheffield: ScHARR, University of Sheffield, 2008.
46. Rabinovich L, Brutscher P-B, de Vries H, Thyssen J, Clift J, Reding A. The affordability of alcoholic beverages in the European Union. Understanding the link between the alcohol affordability, consumption and harms. Technical report. Cambridge: RAND Europe, 2009.
47. Leon D, McCambridge J. Liver cirrhosis mortality rates in Britain from 1950 to 2002: an analysis of routine data. ***The Lancet***;367(95-4):52-6.
48. Marmot M. Evidence-based policy or policy-based evidence? ***British Medical Journal*** 2004;328:906-7.
49. Groves T. Preventing alcohol related harm to health: clamp down on advertising and set a minimum price. ***British Medical Journal*** 2010;340:161-2.
50. Room R, Babor T, Rehm J. Alcohol and public health. ***The Lancet*** 2005(365):519-30.
51. Herttua K, Makela P, Martikainen P. Changes in alcohol-related mortality and its socioeconomic differences after a large reduction in alcohol prices: a natural experiment based on register data. ***American Journal of Epidemiology*** 2008;168(10):1110-1118.
52. Herttua K, Makela P, Martikainen P. An evaluation of the impact of a large reduction in alcohol prices on alcohol-related and all-cause mortality: time series analysis of a population-based natural experiment. ***International Journal of Epidemiology*** 2009;Published online December 7, 2009.
53. Helakorpi S, Makela P, Uutela A. Alcohol consumption before and after a significant reduction in alcohol prices in 2004 in Finland: were the effects different across different subgroups. ***Alcohol and alcoholism*** 2010;45(3):286-92.
54. Ogilvie D, Craig P, Griffin SJ, Macintyre S, Wareham N. A translational framework for public health research. ***BMC Public Health*** 2009;9:116.
55. Ogilvie D, Foster C, Rothnie H, Cavill N, Hamilton V, Fitzsimons C, et al. Interventions to promote walking: systematic review. ***British Medical Journal*** 2007;334:1204-7.
56. Campostrini S, Holtzman D, McQueen D, Boaretto E. Evaluating the effectiveness of health promotion policy: changes in the law on drinking and driving in California. ***Health Promotion International*** 2006;21(2):130-5.
57. Mackay D, Haw S, Pell J. Smoke-free legislation and hospitalizations for childhood asthma. ***New England Journal of Medicine*** 2010;363:34-40.
58. Glasziou P, Chalmers I, Rawlins M, McCulloch P. When are randomised trials unnecessary? Picking signal from noise. ***British Medical Journal*** 2007;334:349-51.
59. Stamatakis E, Ekelund U, Wareham N. Temporal trends in physical activity in England: the Health Survey for England 1991 to 2004. ***Preventative Medicine*** 2007;45:416-23.
60. Blair PS, Sidebotham P, Berry PJ, Evans M, Fleming PJ. Major epidemiological changes in sudden infant death syndrome: a 20-year population-based study in the UK. ***The Lancet*** 2006;367:314-9.
61. Craig P, Dieppe P, Macintyre S, Michie S, Nazareth I, Petticrew M. Developing and evaluating complex interventions: the new Medical Research Council guidance. ***British Medical Journal*** 2008;337:a1655.

62. Ogilvie D, Bull F, Powell J, Cooper A, Brand C, Mutrie N, et al. An applied ecological framework for evaluating infrastructure to promote walking and cycling: the iConnect study. *American Journal of Public Health* 2011;101(3):473-81.
63. Pawson R, Tilley N. *Realistic evaluation*. London: Sage, 2007.
64. Morris J. Uses of epidemiology. *British Medical Journal* 1955;2:395-401.
65. Scottish Government. Clearing the air - smoke-free legislation - latest situation - national compliance data. www.clearingtheairscotland.com/latest/index.html. Last accessed, 7 October 2010.
66. Semple S, Creely K, Naji A, et al. Second hand smoke levels in Scottish pubs: the effect of smoke-free legislation. *Tobacco Control* 2007;16:127-32.
67. Semple S, MacCalman L, Atherton Naji A, et al. Bar workers' exposure to second-hand smoke: The effect of Scottish smoke-free legislation on occupational exposure. *Annals of Occupational Hygiene* 2007;51:571-80.
68. Haw S, Gruer L. Changes in adult exposure to second hand smoke following implementation of smoke-free legislation in Scotland. *British Medical Journal* 2007;335:549-52.
69. Akhtar PC, Currie DB, Currie CE, et al. Changes in child exposure to environmental tobacco smoke (CHETS) study after implementation of smoke-free legislation in Scotland: national cross sectional survey. *British Medical Journal* 2007;35:545-9.
70. Ayres J, Semple S, MacCalman L, et al. Bar workers' Health and Environmental Tobacco Smoke Exposure (BHETSE): Symptomatic improvement in bar staff following smoke-free legislation in Scotland. *Occupational & Environmental Medicine* 2009;66:339-46.
71. Pell J, Haw SJ, Cobbe S, Newby DE, Pell AC, Fischbacher C, et al. Smokefree legislation and hospitalisations for acute coronary syndrome. *Journal of the American Medical Association* 2008;359:482-91.
72. Hyland A, et al. The Impact of Smoke-free Legislation in Scotland: Results from the Scottish International Tobacco Policy Evaluation Project *European Journal of Public Health* 2009;19:198-205.
73. Mackay D, Haw S, Pell J. Impact of smoke-free legislation on prescriptions for nicotine replacement therapy and adult smoking prevalence. For submission
74. Lewis S, Haw S, McNeill A. The impact of the 2006 Scottish Smoke-Free Legislation on sales of nicotine replacement therapy. *Nicotine and Tobacco Research* 2008;10:1789-92.
75. Committee on Seconhand Smoke Exposure and Acute Coronary Events. Secondhand smoke exposure and cardiovascular effects: making sense of the evidence. Washington, DC: Institute of Medicine, 2010.
76. International Agency for Research on Cancer. *Evaluating the effectiveness of smoke-free policies*. Lyon: IARC, 2009.
77. Murphy S, Raisenan L, Moore G, Tudor Edwards R, Linck p, Williams N, et al. A pragmatic randomised controlled trial of the Welsh National Exercise Referral Scheme: protocol for trial and integrated economic and process evaluation. BMC Public Health 2010;10:352.
78. Rutter M. Is Sure Start an effective preventive intervention? *Child and Adolescent Mental Health* 2006;11(6):135-41.
79. Hutchings J, Gardner F, Bywater T, Daley D, Whitaker C, Jones K, et al. Parenting intervention in Sure Start services for children at risk of developing conduct disorder: pragmatic randomised controlled trial. *British Medical Journal* 2007;334:678-82.
80. Belsky J, Melhuish E, Barnes J, Leyland AH, Romaniuk H, National Evaluation of Sure Start Research Team. Effects of Sure Start local programmes on children and families: early findings from a quasi-experimental, cross sectional study. *British Medical Journal* 2006;332:1476-81.
81. Edwards RT, O Ceilleachair A, Bywater T, Hughes DA, Hutchings J. Parenting programme for parents of children at risk of developing conduct disorder. Cost effectiveness analysis. *British Medical Journal* 2007;334:682-7.
82. Brown CA, Lilford RJ. The stepped wedge trial design: a systematic review. *BMC Medical Research Methodology* 2006;6:54-63.
83. Sefton T, Byford S, McDaid D, Hills J, Knapp M. *Making the most of it: economic evaluation in the social welfare field*. York: Joseph Rowntree Foundation, 2002.
84. Fenwick E, Claxton K, Sculpher M. The value of information and the value of implementation: combined and uneven development. *Medical Decision Making* 2008;28(1):21-32.
85. Lorgelly P, Lawson KD, Fenwick EA, Briggs AH. Outcome measurement in economic evaluations of public health interventions: a role for the capability approach. *International Journal of Environmental Research and Public Health* 2010;7:2274-2289.
86. Weatherly H, Drummond M, Claxton K, Cookson R, Ferguson B, Godfrey C. Methods for assessing the cost-effectiveness of public health interventions: key challenges and recommendations. *Health Policy* 2009;93:85-92.
87. Byford S, Mcdaid D, Sefton T. Because it's worth it. *A practical guide to conducting economic evaluations in the social welfare field*. York: Joseph Rowntree Foundation, 2002.
88. Hennekens CH, Buring JE. *Epidemiology in Medicine*. Philadelphia: Lippincott, Williams and Wilkins, 1987.
89. Deeks J, Dinnes J, D'Amico R, Sowden A, Sakarovich C, Song F, et al. Evaluating non-randomised intervention studies. *Health Technology Assessment* 2003;7(27).
90. Meyer BD. Natural and quasi-experiments in economics. *Journal of Business and Economic Studies* 1995;13(2):151-61.
91. Mackenzie M, Catherine OD, Halliday E, Sridharan S, Platt S. Do health improvement programmes fit with MRC guidance on evaluating complex interventions. *British Medical Journal* 2010;340.
92. Waters E, Priest N, Armstrong R, Oliver S, Baker P, McQueen D, et al. The role of a prospective public health intervention study register in building public health evidence: proposal for content and use. *Journal of Public Health* 2007;29(3):322-7.
93. Ukoumunne OC, Gulliford MC, Chinn S, Sterne JAC, Burney PGJ. Methods for evaluating area-wide and organisation-based interventions in health and health care: a systematic review. *Health Technology Assessment* 1999;3(5).

94. Blundell R, Costa Dias M. Alternative approaches to evaluation in empirical microeconomics. ***Cemmap working paper***, 2002.
95. Imbens GW, Wooldridge JM. Recent developments in the econometrics of programme evaluation. ***Journal of Economic Literature*** 2009;47(1):5-86.
96. Lagarde M, Haines A, Palmer N. Conditional cash transfers for improving uptake of health interventions in low and middle-income countries: a systematic review. ***Jounal of the American Medical Association*** 2007;298(16):1900-1910.
97. Lim SS, Dandona L, Hoisington JA, James SL, Hogan MC, Gakidou E. India's Janani Suraksha Yonana: a conditional cash transfer scheme to increase births in health facilities: an impact evaluation. ***The Lancet*** 2010;375:2009-23.
98. Ho D, Imai K, King G, Stuart E. Matchit: non-parametric preprocessing for parametric causal inference. Version 2.4-13. http://gking.harvard.edu/matchit/. Accessed 6 July 2010.
99. D'Agostino R. Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. ***Statistics in Medicine*** 1998;17(19):2265-81.
100. Dusheiko M, Gravelle H, Jacobs R, Smith P. The effect of budgets on doctor behaviour: evidence from a natural experiment. ***Discussion papers in economics***. York: University of York, 2003.
101. Lawlor DA, Harbord RM, Sterne JAC, Timpson N, Davey Smith G. Mendelian randomisation: using genes as instruments for making causal inferences. ***Statistics in Medicine*** 2007;27:1133-1163.
102. Brookhart MA, Rassen JA, Schneeweiss S. Instrumental variable methods in comparative safety and effectiveness research. ***Pharmacoepidemiology and Drug Safety*** 2010;19:537-54.
103. McClellan M, McNeil BJ, Newhouse JP. Does more intensive treatment of acute myocardial infarction in the elderly reduce mortality? Analysis using instrumental variables. ***Journal of the American Medical Association*** 1994;272(11):859-66.
104. Stukel TA, Fisher ES, Wennberg DE, Alter DA, Gottlieb DJ, Vermuelen MJ. 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. ***Jounal of the American Medical Association*** 2007;297:278-85.
105. Hernan M, Robins JA. Instruments for causal inference. An epidemiologist's dream? ***Epidemiology*** 2006;17(4):360-72.
106. Imbens GW, Lemieux T. Regression discontinuity designs: a guide to practice. ***Journal of econometrics*** 2008;142(2):615-35.
107. Ludwig J, Miller DL. Does Head Start improve children's life chances? Evidence from a regression discontinuity design. ***Quarterly Journal of Economics*** 2007:159-208.
108. Neidell M. Air quality warnings and outdoor activities: evidence from Southern California using a regression discontinuity design. ***Journal of Epidemiology and Community Health*** 2009;epub ahead of print.
109. Belot M, James J. Healthy school meals and educational outcomes. ***ISER Working Paper Series***, 2009.
110. Groenwald RH, Nelson D, B, Nichol KL, Hoes AW, Hak E. Sensitivity analyses to estimate the potential impact of unmeasured confounding in causal research. ***International Journal of Epidemiology*** 2010;39:107-117.
111. von Elm E, Altman D, Egger M, Pocock SJ, Gotzsche P, Vandenbroucke JP, et al. Strengthening the reporting of observational studies in epidemiology (STROBE) statement: guidelines for reporting observational studies. ***British Medical Journal*** 2007;335:335-8.
112. DesJarlais DC, Lyles C. Improving the reporting quality of nonrandomized evaluations of behavioural and public health interventions: the TREND statement. ***American Journal of Public Health*** 2004;94(3):361-6.
113. Fox MP. Creating a demand for bias analysis in epidemiological research. Journal of Epidemiology and Community Health 2009;63:91.
114. Jurek AM, Maldonado G, Spector LG, Ross JA. Periconceptional maternal vitamin supplementation and childhood leukaemia: an uncertainty analysis. ***Journal of Epidemiology and Community Health*** 2009;63:168-72.
115. Turner R, Spiegelhalter D, Smith G, Thompson S. Bias modelling in evidence synthesis. ***Jounal of the Royal Statistical Society (Series A Statistics in Society)*** 2009;172:21-47.
116. Petticrew M, Roberts H. ***Systematic reviews in the social sciences. A practical guide***. Oxford: Blackwell, 2006.
117. Reeves B, Deeks J, Higgins JP, Wells GA. Including non-randomised studies. In: Higgins JP, editor. ***Cochrane Handbook for Systematic Reviews of Interventions***. 5.0.1 ed: The Cochrane Collaboration, 2008.
118. Armstrong R, Waters E, Jackson N, Oliver S, Popay J, Shepherd J, et al. ***Guidelines for systematic reviews of health promotion and public health interventions***. Version 2. Melbourne: Melbourne University, 2007.
119. Wells G, Shea B, O'Connell D, Peterson J, Welch V, Losos M, et al. The Newcastle-Ottawa Scale (NOS) for assessing the quality of nonrandomised studies in meta-analyses. http://www.ohri.ca/programs/clinical_epidemiology/oxford.htm. Accessed 6 July 2010.
120. Diamond J, Robinson JA. Prologue. In: Diamond J, Robinson JA, editors. ***Natural Experiments of History***. Cambridge, Mass: The Belknap Press, 2010.
121. Sims M, Maxwell R, Bauld L, Gilmore A. Short term impact of smoke-free legislation in England: retrospective analysis of hospital admissions for myocardial infarction. ***British Medical Journal*** 2010;340:c2161.

Annex 1
Alternative definitions of natural experiments

'Outcomes are compared across treatment and control groups, and both a priori reasoning and empirical evidence are used to validate the assertion of randomisation. Thus, random or "as if" random assignment to treatment and control conditions constitutes the defining feature of a natural experiment.'

- Dunning 2008[7]

'[N]atural experiments as opposed to random experiments imply acts of nature, or more generally, exogenous interventions demarcating observations in theoretically important ways. However, the key distinction is that the assignment mechanism is out of the control of the researcher, whereas in a controlled experiment the assignment mechanism is generated by the researcher for the experiment itself. In a natural experiment, some external force intervenes and creates comparable treatment groups in a seemingly random fashion.'

- Robinson et al. 2009[4]

'Good natural experiments are studies in which there is a transparent exogenous source of variation in the explanatory variables that determine the treatment assignment.'

- Meyer 1995[90]

'[T]he so-called natural experiment ... typically considers the policy reform itself as an experiment and tries to find a naturally occurring control group that can mimic the properties of the control group in the properly designed experimental context.'

- Blundell and Costa Dias 2002[94]

'A natural experiment constitutes some circumstance that pulls apart variables that ordinarily go together and, by so doing, provides some sort of equivalent of the manipulations possible in an experiment deliberately undertaken by a researcher.'

- Academy of Medical Sciences 2007[26]

'Naturally occurring circumstances in which subsets of the population have different levels of exposure to a supposed causal factor, in a situation resembling an actual experiment where human subjects would be randomly allocated to groups.'

- Last 1995[15]

'The term natural experiment describes a naturally-occurring contrast between a treatment and a comparison condition. Often the treatments are not even potentially manipulable[.]'

- Shadish et al. 2002[19]

'A technique that frequently proves fruitful in [the] historical disciplines is the so-called natural experiment or the comparative method. This approach consists of comparing – preferably quantitatively and aided by statistical analyses – different systems that are similar in many respects but that differ with respect to the factors whose influence one wishes to study.'

- Diamond and Robinson 2010[120]