The Academy of Medical Sciences

Identifying the environmental causes of disease: how should we decide what to believe and when to take action?

An Academy of Medical Sciences working group report chaired by Sir Michael Rutter CBE FRS FBA FMedSci

November 2007

The Academy of Medical Sciences

The Academy of Medical Sciences promotes advances in medical science and campaigns to ensure these are converted into healthcare benefits for society. Our Fellows are the UK's leading medical scientists from hospitals and general practice, academia, industry and the public service.

The Academy plays a pivotal role in determining the future of medical science in the UK, and the benefits that society will enjoy in years to come. We champion the UK's strengths in medical science, including the unique opportunities for research afforded by the NHS, encourage the implementation of new ideas and solutions – often through novel partnerships, promote careers and capacity building and help to remove barriers to progress.

ISBN No: 1-903401-16-X

The Academy of Medical Sciences

Identifying the environmental causes of disease: how should we decide what to believe and when to take action?

An Academy of Medical Sciences working group report chaired by Sir Michael Rutter CBE FRS FBA FMedSci

November 2007

Acknowledgements
The Academy of Medical Sciences is most grateful to Professor Sir Michael Rutter CBE FRS FBA FMedSci and the members of the working group for undertaking this study. The Academy wishes to thank the review group, the Academy's Officers, Council and staff, participants at the workshop and all respondents to the consultation for their informative comments and support. The Academy is grateful to the University Hospitals Association for its support.

Disclaimer
This report is published by the Academy of Medical Sciences and has been endorsed by its Officers and Council. Contributions by the working group and respondents to the call for evidence are made purely in an advisory capacity. The review group added a further 'peer-review' stage of quality control to the process of report production. The reviewers were not asked to endorse the report or its findings.

The members of the working group and the review group participated in this report in an individual capacity and not as representatives of, or on behalf of, their affiliated hospitals, universities, organisations or associations. Their participation should not be taken as endorsement by these bodies.

Contents

Summary

1. Scarcely a day goes by without some new report of a study claiming to have discovered a new important environmental cause of disease. Often these concern serious disorders such as cancer or heart disease and sometimes they implicate factors such as toxins or diet that are readily susceptible to modification. The problem is that few of these findings are confirmed by subsequent research and, occasionally, new studies even find the opposite. If many of these causal claims turn out to be mistaken, how should we decide what to believe and when to take action?

2. The challenge for the working party was to consider the types of research needed to identify environmental causes of disease when, for practical or ethical reasons, they could not be experimentally investigated. Inevitably, therefore, our attention had to be focused on non-experimental studies observing associations between specific risk features and different disease outcomes. We considered the strengths and limitations of such non-experimental studies and what steps can be taken to reduce the uncertainties about their supposed causal effects.

3. In order to build the rich evidence base that underpins the conclusions of this project we issued a call for evidence to which over 70 written submissions were received. This was buttressed by the findings of a successful workshop that brought together a wide range of stakeholders. The evidence obtained from these two activities was considered alongside many published papers. The final report was subject to peer-review.

4. We started our deliberations by asking whether there was good reason to suppose that environmental features were likely to play an important role in the causation of disease. We found that the evidence is clear cut; environmental influences are both strong and important in the causal processes leading to most common diseases. Nevertheless, knowledge on the specifics of these environmental influences, and of the biological pathways through which they exert their causal effects, is decidedly limited. We concluded that priority needs to be given to high quality research using designs that could help identify the environmental components of the causal pathways that lead to disease.

5. Sometimes people have wanted research to identify the single basic cause of disease. We concluded that this was not the right question. Most common diseases involve the coming together of multiple environmental and multiple genetic causes. Accordingly, the question needed to be: how can we identify whether some specific environmental factor has a true causal effect that contributed to the development of a disease – meaning that, if it were not present, the rate of that disease would be less? The implication is that knowledge about the causes of disease can have an important impact on its treatment, diagnosis or prevention.

6. We concluded that non-experimental methods are fundamental to clinical practice and policymaking. Provided stringent criteria are met, non-experimental research can, and does, give rise to valid inferences on the environmental causes of disease. This has important implications for both public policy and the treatment of individual patients.

7. The examples of non-experimental research that have played a key role in the effective identification of environmental causes of disease point to the importance of integrating findings across a range of research strategies – experimental and non-experimental, in humans and non-humans. With very few exceptions, no one research approach, and no one study, provides conclusive evidence. Moreover, the testing of causal inferences usually involves testing in several different populations and several different contexts to determine how far conclusions can be

generalised and to assess whether effects are specific to a given context. The totality of evidence from all sources should be brought together in order to reach sound conclusions.

8. In clinical medicine, randomised controlled trials (RCTs) have rightly become the preferred method for testing the efficacy of treatments. But there are many possible environmental causes of disease that could not be investigated using RCTs because they would be impractical or unethical in humans. Nevertheless, we affirm the immense strength of RCTs that derives from the combination of a controlled application of some planned intervention, plus randomisation that ensures that features that could affect the outcome in an unplanned way are randomly distributed between the groups to be compared.

9. There are circumstances in which RCTs could be applied in relation to protective factors or to causes of disease that derive out of risks from therapeutic interventions. We encourage researchers, policy makers and funders to make greater use of RCTs in these uncommon circumstances. There are also considerable advantages in combining RCTs with non-experimental studies, because each has different patterns of strengths and limitations.

10. Since, for most possible environmental causes, RCTs are not feasible or ethical, we focused most of our attention on the use on non-experimental research designs of various kinds. We draw attention to the value of natural experiments that, by pulling apart variables that ordinarily go together, can provide much needed additional research leverage. Sometimes, this is because of their power to differentiate between genetic and environmental causal effects, and sometimes because of their power to avoid the biasing effect of social selection resulting from individual choice or behaviour. We recommend that greater consideration be given to their use.

11. All research findings will be affected by chance variation, and random error, but there are well established statistical techniques for taking these into account. In non-experimental research, by contrast with controlled experiments, there is an additional concern about unidentified systematic error creating a bias that may lead to a misleading causal inference. In our report, we discuss some of the design features and statistical approaches that can help in minimising this problem. All of these, however, are prone to the error created by sources of bias that were not conceptualised and therefore were not measured. The best protection is provided by the use of background knowledge to enable the formulation and testing of hypotheses on possible alternative explanations for the observed associations.

12. Over 40 years ago, the renowned statistician Sir Austin Bradford Hill FRS set out an influential set of guidelines to help decide when a statistical association was likely to reflect true causation. Since then, there have been major research developments and, hence, there was a need to re-examine his guidelines. We concluded that the guidelines have stood the test of time and remain useful. He emphasised the need for multiple criteria, and we agree. Also, he stressed the need to test competing non-causal explanations and, again, we agree. He argued the importance of biological plausibility as one guide. We, too, point to the important contributory role of experimental evidence showing a biological mechanism likely to account for the causal effect. Similarly, we urge caution before accepting a causal claim if no plausible way in which the putative causative factor could operate can be suggested. Knowledge on biological mechanisms is very dependent on experimental evidence and we note the value of animal models in that connection. We also note that what is ill understood at one time may become better understood at a later time in the light of other scientific advances. The one criterion that does need some modification is diagnostic specificity, in which a single cause

has an effect on only one disease, in view of the extensive evidence that some causal agents affect multiple diseases. However, to some extent, this arises because some agents involve risks that operate through several different causal mechanisms. The ill health consequences of smoking well illustrate that point. We conclude that specificity does help if it is strong (as it is with some of the prenatal teratogenic effects) but lack of specificity should not rule out causation.

13. The evidence that we have reviewed shows that most misleading causal claims (whether from experimental or non-experimental research) stem from poor quality studies of small biased samples, often reported in conferences and not subjected to rigorous review by fellow researchers in scientific journals. We urge particular caution in the reporting of such studies, we emphasise the ever present need for quality in research, and we reiterate the need for all new findings to be subjected to testing by independent investigators on different samples (i.e. replicated) and for new evidence to be evaluated in the light of existing knowledge and other research findings.

14. When attention is confined to high quality non-experimental research, most has given rise to findings that are confirmed by other research designs. Exceptions mainly involve situations in which there is a substantial likelihood of bias stemming from people's actions in choosing or shaping their exposure to risky or protective environments. When such biases are expected, there is an especially great need to use designs that may be able to take account of such allocation biases. That is where the combination of non-experimental studies with experimental approaches (including both basic science and RCTs) in humans and other animals is particularly important.

15. All research is provisional in the sense that it solves some problems and opens up many new questions. Even the best research may need to be reinterpreted if later scientific advances or new evidence cast doubt on the interpretations or meaning of findings. This means that all causal inferences involve some degree of uncertainty. Nevertheless, both individual medical practitioners and policymakers have to make decisions based on whatever evidence is available at the time. The decision not to act is as real and important a decision as one that leads to intervention or policy change. Such decisions require careful judgments on the balance of likely risks and benefits associated with different decisions.

16. Claims about the identification of a cause of a serious disease often stimulate great interest among the public because of their relevance for people's daily lives. A common cause for confusion is the failure to differentiate between relative risk (i.e. whether the risk after exposure to some causal factor is greater or lesser than that in the general population) and absolute risk (i.e. the probability that they will actually get the disease in question). Confusion may also arise because researchers and/or the media do not make clear whether the risks apply to everyone or only to some small segment of the population. All of those involved in communicating or acting upon research into causes (whether stemming from experimental or non-experimental studies) should be mindful of its possible impact on people's behaviour and on public policy.

17. Research into the environmental causes of disease will only prove useful if careful attention is paid to how the research is generated, interpreted, communicated and acted upon. In view of the multiple strategic and technical issues involved in the identification of environmental causes, especially in their study through non-experimental methods, it is crucial to integrate science into policy making. Research into the causes of disease is so important to people's lives, we recommend that steps should be taken to actively involve the public and patient organisations by inviting them to participate in the scientific advisory committees of funding bodies. Similarly,

because well conducted non-experimental research is so important in the identification of environmental causes of disease, we urge that funders make clear their support.

18. Policymakers often have to make public health decisions rapidly using existing research evidence, rather than waiting for further research to be generated and completed. It is therefore necessary to integrate vigorous piloting into the implementation of new policies or practice. Because even well based policy changes may not bring about the expected benefits, it is also crucial that the changes be introduced in a manner that allows rigorous evaluation, and that funds be provided for such evaluation.

19. The challenges inherent in the interpretation of high quality non-experimental and experimental evidence concerning the identification of environmental causes of disease means that everyone has a responsibility to deal with findings in a considered and balanced fashion. We recommend that consideration should be given to the possibility of making accurate communication of results a requisite of funding.

20. It is evident that change cannot only come from above. Accordingly we have provided sets of guidelines tailored to the roles of different stakeholder groups – ranging across the continuum from the undertaking of research to its translation into policies and practice.

Recommendations

Our recommendations apply to all kinds of medical science that is concerned with the identification of environmental causes of disease. Our focus, however, is particularly on non-experimental studies, because they involve the special features of uncertainty on causal inference and are of major public health importance.

Recommendation 1
Government should build upon their recent efforts to integrate science into policy making by further increasing capacity building by means of:

- **Embedding researchers into policy teams.**
- **Providing senior civil servants with scientific training.**
- **Seconding scientists to government.**
- **Building a cadre of 'evidence brokers' within government who are trained in both science and policy.**

Recommendation 2
The Research Base Funders' Forum should lead an initiative to reaffirm funders' support, where appropriate, for high quality non-experimental research into the environmental causes of disease, encourage studies to test previous findings in different circumstances, and undertake systematic reviews.

Recommendation 3
The Department of Health and other relevant government departments should ensure that there is a greater emphasis on both pilot studies and systematic rigorous evaluations of the effects of interventions in developing and implementing health policy.

Recommendation 4
The Research Base Funders' Forum should lead an initiative to foster responsibility for the accurate communication of non-experimental research. This should include consideration of whether it would be feasible to make accurate communication of results a requisite of funding.

Recommendation 5
The Department of Health, Research Councils, and charities funding research into the environmental causes of disease and interventions to prevent or treat disease should continue to involve the public and patient organisations by inviting them to participate in their expert scientific advisory committees.

Guidelines

In addition to the recommendations for action, we have prepared more detailed guidelines to assist those involved in the production, interpretation, communication and implementation of research findings on the environmental causes of disease. The several guidelines are organised in terms of issues that are particularly relevant to different groups of people.

Guidelines for researchers

1. Consider the relative merits and limitations of different research designs that may be possible for the postulated environmental cause being studied.
2. Plan the research to obtain the most comparable groups possible and stratify when appropriate to increase comparability.
3. Think carefully about possible confounders, especially those creating selection bias, indication bias, or ascertainment bias. Make sure that all of these are measured as carefully, accurately and systematically as possible.
4. Pre-plan and record the planned analyses in order to avoid later data dredging. It is desirable to have a protocol for all types of research – both experimental and non-experimental.
5. In exploratory non-experimental research, it is often necessary to modify analytic plans iteratively in the light of earlier analytic findings. When that is the case, particular care should be taken to examine alternative interpretations and to keep secondary analyses conceptually separate from initial plans.
6. Whenever possible, use natural experiments or employ an RCT when that is feasible and ethical.
7. Make sure that the study is adequately statistically powered.
8. Whenever possible, build in replications across different samples. Consider also the possible value of meta-analyses or other methods of combining samples.
9. Consider, and systematically test for, alternative non-causal explanations for all findings. Do not be satisfied until rigorous exploration of these alternatives has been undertaken and shown not to account for the causal inference.
10. Use rigorous methods of analysis to test for the effects of confounders (especially of the types noted above) and include sensitivity analyses in what is done.
11. Whenever possible, test for the mediating mechanisms involved in causal pathways. Often this will require the bringing together of different research strategies.
12. Report the findings with careful attention to other research (noting when findings differ from those now being reported) and to what is known on biological mechanisms.
13. Make explicitly clear the extent to which findings are likely to warrant a causal inference and warrant generalisation to other samples. Do not be tempted to make claims that cannot be adequately justified.
14. Resist any pressures from the funding agency, from your employing authority, or the media to exaggerate the claims. Strenuously resist pressures from any sources to censor, distort or bias your report of findings.
15. Do not persist with causal claims when new evidence indicates they were mistaken or should be overturned by the findings of more statistically powerful studies.
16. Consider the value of constructively critical reviews of key topics.
17. When talking or writing for a broad audience, be mindful of the need to express concepts and findings clearly in an understandable fashion. Recognise the value of public engagement and appreciate the importance of doing this very well.
18. Disclose all conflicts of interest.

Guidelines for editors of science and medical journals

1. Be more vigilant when considering for publication potentially controversial findings about the causes of disease, or reports based on non-experimental methods that might conceivably influence clinical practice or change health behaviours.
2. Every editor and peer-reviewer should consider himself or herself to be a guardian of research integrity and public trust in science.
3. Strengthen the collective responsibility of co-authors who need to take a shared ownership of the totality of a research study and the messages it might impart.
4. Support the creation of reporting guidelines for research into the causes of disease, particularly using non-experimental methods, and, when such guidelines are available and approved by reputable scientific institutions, apply those recommendations to the papers submitted for publication.
5. At the time of publication of any high risk paper on the environmental causes of disease, the editor should consider running an accompanying editorial or critique to place that work in context of the totality of available evidence, with particular reference to public health issues.
6. Any press materials issued by the journal or by the host institution or by the funder of work to investigate the environmental causes of disease should consider the way that research may be reported by the media. If there is any risk of harm to public health, individual health behaviours, or clinical practice, the journal/institution/funder should act appropriately to limit that harm.
7. In preparing press releases, care should be taken to ensure that the level of absolute risk is both provided and explained.

Guidelines for science or medical writers and journalists

1. Pay detailed attention to the methodology of all studies being reported. Important questions to consider include:
   - What was the sample?
   - What were the measures?
   - How strong were the effects in both relative and absolute terms?
   - Has there been adequate attention to alternative explanations, and to good control of possible confounding variables?
   - Has the finding been replicated?
   - Is there supporting experimental or quasi-experimental evidence?
   - Are the findings in keeping with what is known about disease mechanisms?
2. Whilst it may not be appropriate to offer extensive discussion of all these details when writing or speaking to the general public, key aspects can be communicated successfully using clear, jargon free, language.
3. The science or medical correspondent needs to have an appropriate grasp of the scientific issues in order to know how best to convey what was novel, interesting and important in the research.
4. Exercise appropriate judgment in identifying and drawing attention to those points of design that are particularly relevant to the study in question – especially when ignoring them might lead to misunderstanding.
5. Bear in mind the research track record of the researchers and of their employing institutions.
6. Consider whether there are any conflicts of interest that might lead to possible bias.
7. Seek to determine the theory or set of biological findings that constitute the basis for the research - noting how this fits in with, or forces changes in, what we already know or believe.
8. Whilst paying appropriate attention to competing views, be wary of creating spurious and misleading 'balance' by giving equal weight to solid research evidence and weakly supported idiosyncratic views.
9. Be very wary of drawing conclusions on the basis of any single study, whatever its quality.
10. When considering public policy implications, draw a careful distinction between relative risk (i.e. the increased probability of some outcome given the disease causing factor) and absolute risk (i.e. the probability of that disease outcome in those with the disease risk).
11. Use simple counts to describe risk whenever possible, rather than probabilities.
12. Be careful, insofar as the evidence allows, to clarify whether the causal effect applies to everyone or only to a small special sub-segment of the population.
13. Set the causal factor you are describing in the context of all known causal factors, whilst explaining that there may be others, as yet unknown or unsuspected.
14. In writing about research, seek to educate and engage readers with the science and to encourage them to think critically.

Guidelines for policymakers

1. Make sure that your scientific advisors give you a clear and balanced assessment of the strength of the scientific evidence and of its potential for generalisation.
2. Be especially careful to be critical of evidence that supports existing policy. Consider carefully when contrary evidence warrants a change in policy. But, also consider when further research is needed (it usually will be) to assess the effects (both beneficial and harmful) of the change in policy.
3. At all times, keep absolutely clear the distinction between a lack of research to show something, and positive replicated research that shows some causal claim to be false (or highly uncertain). The two are quite different and confusion between the two is especially likely to lead to the need for a later withdrawal of assurances that all is well.
4. There are very few (if any) interventions that do not carry a degree of risk for some individuals even though there may be major benefits for the population as a whole. The balance between the two will always need to be carefully assessed. Consider carefully the trade-off between the likelihood of benefits and risks of preventive or therapeutic interventions and the possible harm associated with no action.
5. Policy decisions will always need to be based on the particular circumstances in which interventions are likely to make an impact, and not just on the evidence on which interventions are most effective in optimal circumstances.
6. It is the duty and responsibility of policymakers (at both a national and local level) to make value judgments, but it is also crucial to pay attention to research findings on which interventions are likely to be most effective in achieving the desired aims.
7. Take any necessary actions to make sure that the raw data of government funded research are available for scrutiny and, if appropriate, for further analyses.
8. Many policy decisions based on research causes will cut across different government departments. Ensure effective interdepartmental communication and consultation.
9. Consider the value of longer-term research that goes beyond the term of the current government.
10. Use pilot studies to assess the implementation of new policies in order to plan a more definitive evaluation of their effects.
11. Consider the value of strengthening the links between policymakers and researchers.
12. Build on the iterative nature of well functioning interactions between policymakers and researchers.
13. Involve patient organisations and the general public in the decision making that may result from research into the causes of disease.

Guidelines for clinicians and healthcare practitioners

1. Clinicians, like policymakers, have to act on the evidence available at the time when deciding how to advise or treat each patient. There is not the luxury of waiting for uncertainty to be reduced to a minimal level. Bear in mind that 'not to act' is just as much an action as providing an intervention.
2. Be alert to the need to think and act differently as a consequence of a new set of research findings, but be wary of claims made on the basis of a single study, or claims that one method is best in all circumstances. Most well based understanding of causal processes comes from the combined results of different types of research.
3. Use continuing professional development as a way of keeping up with clinically relevant research advances and new possibilities of conceptualisation. Use the experience of reading journals, or attending seminars/lectures, or clinical teaching occasions, as a means of gaining a better understanding of the strengths and limitations of different forms of research.
4. When patients ask about some new claim in the media, be prepared to discuss the claim openly, but, if necessary, ask for time to inform yourself better about the study leading to the claim.

Guidelines for funders

1. Recognise the crucial importance of identifying the environmental causes of disease and protective influences against them.
2. Appreciate that non-experimental research has a crucial part to play in this search for causes because so few causes are susceptible to straightforward ethical experimental manipulation in humans.
3. Differentiate between purely descriptive non-experimental studies and those that carry the potential to identify causes.
4. Recognise the value of 'natural experiments' and do not dismiss them on the grounds that they deal with unusual samples since they have to do so.
5. Recognise the value of creative new research strategies and do not, when funding is tight, retreat to a position of conservatism.
6. Be willing to support 'replication' studies to test hypotheses thrown up by exploratory investigations.
7. Be willing to support critically constructive systematic reviews.
8. Provide incentives for researchers to communicate accurately the results of studies to identify the environmental causes of disease and to put their findings in the wider context of other research.
9. Any press materials issued by the funder, host institution or journal about work to investigate the environmental causes of disease should consider the way that research may be reported by the media. If there is any risk of harm to public health, individual health behaviours, or clinical practice, the journal/institution/funder should act appropriately to limit that harm.

1 Introduction

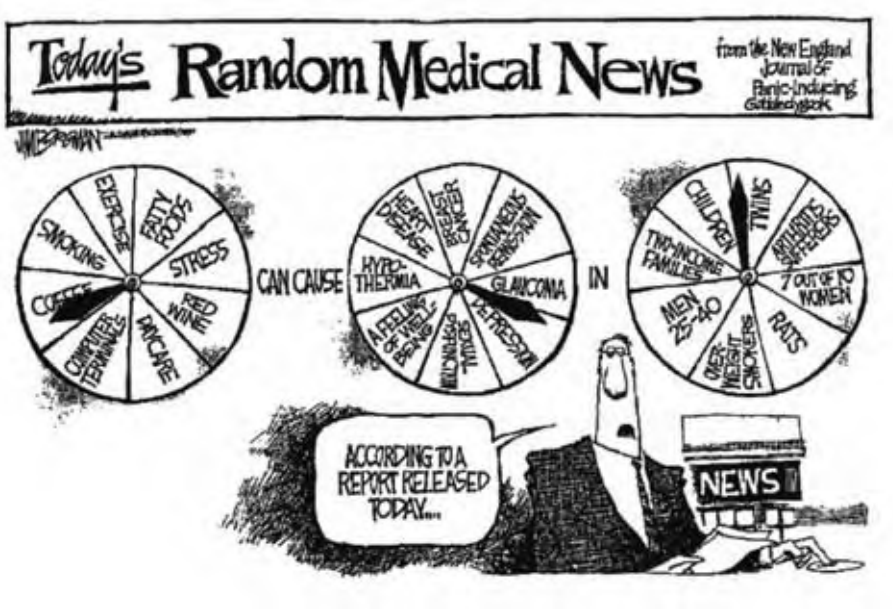

Figure 1: Cartoon that illustrates the confusion that sometimes accompanies research into the causes of disease. Reproduced with kind permission of the New York Times.

At the heart of the matter is the importance of understanding what causes disease. It is a dull week in which there is not a new claim that a disease is caused by some environmental factor or other. As Taubes (1995) commented, these include an astonishing range of supposed disease causing agents – including hair dyes, coffee, a high cholesterol diet, high alcohol mouth washes, pesticides, stress at the work place, mobile phones, eating red meat, and living near overhead power lines. Few of these claims are confirmed by further research, and some studies even find the opposite. The diseases that are claimed to be caused by these features are often serious; thus, many of the claims apply to some form of cancer or heart disease. Moreover, many of the putative disease causing agents can be manipulated. We could choose to alter our diets, and actions could be taken to reduce exposure to toxins or pesticides. But, if many of these claims turn out to be mistaken, how should we decide which findings to believe? Moreover, if there is so much uncertainty, how should policymakers know when and how to take steps to deal with the supposed disease causing agents? These are the key questions that we sought to address in this report.

It is against this background that the Academy of Medical Sciences had become increasingly concerned about the atmosphere of scepticism among both professionals and members of the public regarding claims on the identification of causes of disease. On the one hand, there were solid, well established, examples of research that not only had identified causes of disease, but also had led to important changes in policy and practice, such as the relationship between smoking and lung cancer. On the other hand, there were also numerous examples of claims that proved fallacious such as the purported protective effects of hormone replacement therapy (HRT) on cardiovascular disease. Accordingly, in 2006 the Academy convened a working group to help address these issues. The objective was to produce a set of principles, illustrated with specific examples, that might provide guidelines on when and how to assess causal claims and when to recommend that the causal inference is sufficiently secure to warrant action.

In clinical medicine, as in science more generally, there is a tradition of relying, whenever possible, on the findings of experimental research. The reason for this

reliance is that the most confidence can be placed on a claim that a cause has been identified if experiments have been undertaken to manipulate the causal variable, under controlled conditions, in order to determine if it truly brings about the outcome in question. Following this tradition there has been increasing use of RCTs in medicine in order to determine which treatments are truly effective (or ineffective). Indeed, there have been strong claims that they provide the only acceptable 'gold standard' of proof on causation. The problem, however, is that most of the causal claims concern factors that are not open to manipulation in an ethical manner. Obviously, it would not be acceptable deliberately to expose people to pesticides to find out if they caused disease. Accordingly, the working party mainly focused on studies of possible causes that did not involve experiments. Most of these relied on non-experimental evidence of one kind or another, although increasing use has been made of different forms of 'natural experiments'; and we summarise what is known of their utility. A description of non-experimental methods can be found in Box 1.

Box 1 What do we mean by non-experimental methods?

Throughout this report the term 'non-experimental' refers to the systematic, often quantitative, observation of biomedical phenomena in a population without deliberately planned scientific manipulation (or control) of the variables under investigation. The objective of such research is to identify statistical associations from which causes can be inferred. The techniques most used include prospective cohort studies, case-control comparisons and ecological studies, but the approach extends more widely.

Before examining these studies that rely on correlations or associations between a postulated causal agent and some disease outcome, it was first necessary to consider what is meant by a 'cause', and the different types of studies into causes. Deliberately, we focused only on environmental causes that carried public health implications and which might provide the basis for preventive policies.

We did not consider 'fixed' causes such as genetic variations, or age, or sex, other than to note that some of the evidence on 'fixed' causes also relied on associations and, hence, involved the same sort of inferential problems that we considered. Equally, we did not consider treatment studies other than to discuss the implications of the few instances in which both RCTs and non-experimental studies were applicable.

The working group's terms of reference were as follows:

- To investigate the strengths, limitations and potential of non-experimental methods for the identification of environmental causes of disease.
- To investigate the lessons that might be learnt from successful and less successful examples of non-experimental research.
- To investigate how non-experimental studies should deal with complex multifactorial causes.
- To investigate how experimental and non-experimental approaches should be coordinated to identify causal mechanisms of disease.
- To investigate how non-experimental research is communicated, the value placed by individuals, society and government upon such research and how the results impact on policy and on the decision making of individuals.

It is through these Terms of Reference that the working party sought to address the following questions:

- When are causal inferences from non-experimental studies justifiable?
- Can non-experimental studies give rise to causal inference?
- Can non-experimental studies be misleading?

- Why are there so many conflicting claims about the causes of disease?
- Do RCTs constitute the only satisfactory means of establishing causation?
- Is there a statistical approach that adequately deals with confounding variables?

It was agreed that the working group would not draw conclusions in relation to any particular disease outcome or set of risk factors other than as examples to help illustrate broader principles.

Details of the working group, and review group, are given in Appendix II.

The Academy issued a call for evidence in October 2006, to which over 70 written submissions were received from a wide range of individuals and organisations. The information gathered was analysed and assimilated alongside many published papers.

The Academy held a well attended workshop in June 2007 to seek further views from stakeholders and inform the development of the working group's conclusions. The organisations and individuals who were involved in these activities are listed in Appendix III.

The report is aimed at all those involved in the production, interpretation, communication and implementation of research findings on the environmental causes of disease, particularly:

- Researchers
- Editors of science and medical journals
- Science and medical writers and journalists
- Clinicians
- Policymakers
- Funders

While there is much to be gained from considering the report in its entirety the busy reader may wish to consider 'generic' chapters, such as the introduction and conclusions, then focus on those chapters that are relevant to their specific interests.

2 What is a cause?

2.1 What is meant by a cause when there are multiple causal elements?

Before proceeding with a discussion of how to identify causes, we need to start with a consideration of what we mean by a cause. The need arises because very few diseases or disorders have a single basic cause that is both necessary and sufficient. Perhaps that might apply to infectious diseases. Thus, no-one can get a streptococcal throat infection without exposure to the streptococcus, and the development of the infection does not require the occurrence of other causal influences. Nevertheless, even in this case, it is necessary to recognise the very substantial individual variation in response to the infective agent – a variation that reflects genetic susceptibility, immune status that may be affected by stress, and overall levels of nutrition – to give but three rather different features.

Similarly, Mendelian genetic disorders such as tuberous sclerosis, cystic fibrosis or Huntington's disease represent genetic determinism. No-one can get the disease without the relevant genetic mutation and whether or not they do so does not require the presence of any other genetic or environmental influence. But, again, there is substantial individual variation in the clinical effects, the causes of which often remain ill understood.

Even in these extreme cases, the apparent one to one causal effect constitutes an oversimplification. Rothman and Greenland (1998) used the example of the turning on a light by flicking a switch. It would seem that here there is a direct one to one causal effect because it is not necessary to do anything else to cause the light to go on. However they pointed out, following Mill (1843), that actually the causal effect comprises a constellation of components that act in concert. Thus, the light will not go on unless the bulb is functional, the electric circuit is intact, the required voltage is available etc. Most causes operate in this fashion (see MacMahon *et al.* 1960, who proposed the metaphor of a 'web of causations').

When considering multifactorial medical disorders such as diabetes, coronary artery disease, asthma, schizophrenia or depression, further complications in the causal process have to be appreciated. To begin with, they do not have a single necessary and sufficient cause. That is, they are not caused by a single feature that has to be present and which is sufficient on its own to cause the disease. Rather, they have multiple causal influences each of which, in conjunction with others, contributes to the causal process.

Often, too, there may be several different causal pathways leading to the same end point (Rutter 1998). For example, chronic obstructive pulmonary disease may have its starting point in severe asthma, in heavy smoking, or in lung infection. Each of these starting points is also the end of a prior causal process. Thus, the heavy smoking is likely to involve a genetically influenced susceptibility, the availability of cigarettes, and the operation of social pressures. The fact that the smoking habit persists will also be influenced by the heavily addictive effects of nicotine. Not only is there not just one cause, but none is the basic cause. What is most important will depend on which elements in the causal pathway can be manipulated most successfully. In some circumstances, the focus may need to be on the final stages of the causal process. In other circumstances, the focus may need to be on some much earlier point – perhaps at a stage when intervention might be both feasible and effective.

For all these (and other) reasons, there is no point in seeking to identify 'the' cause of a multifactorial disorder, because there is no such thing. This appreciation led Mackie (1965; 1974) to refer to causes that are 'insufficient

but necessary components of unnecessary but sufficient causes' – INUS in short. What this apparently complicated, but actually very simple, concept means is that the overall causal nexus in its entirety is a sufficient cause of the condition being considered – that is, it is enough to cause the disease without the operation of any other influences. On the other hand, it is an unnecessary cause, because it represents only one out of several causal pathways leading to that outcome. Conversely, the INUS are insufficient because, on their own, they will not cause the disease. They are necessary because, if all other components are held constant, the disease will not occur in their absence. Thus, referring to individual INUS, Rothman and Greenland (2002) defined a cause of a disease occurrence as an antecedent event, condition or characteristic that was necessary (given that other conditions are fixed) for the occurrence of the disease at the moment it occurred. Without that causal influence, the disease would not have developed or would have done so at some later time. That is the central point in deciding what is meant by a cause.

As Robins and Greenland (1989) pointed out, there is the implication that changing a causal factor will actually reduce the population's burden of disease, either by reducing the overall number of cases or by making the disease occur later than it would have done otherwise. In other words, the importance of a causal inference is that it has potentially important implications for prevention or intervention. The authoritative United States Surgeon General's report on the health consequences of smoking (Office of the US Surgeon General 2004) used a comparable concept of a cause, and we continue in the same tradition.

Two other considerations need to be noted before dealing with the identification of causes. First, some causes may operate only in certain contexts. Thus, when there is a strong gene-environment interaction, environmental causes may be operative only when combined with a specific type of genetic susceptibility (Caspi & Moffitt 2006). Alternatively they may operate only in particular age groups – as appears to be the case with the effects of heavy use of cannabis in predisposing to schizophrenia (Arseneault *et al.* 2004; Moore *et al.* 2007; Zammit *et al.* 2002). The evidence suggests that the effect only applies during the years before adulthood. Alternatively, the causal effect may apply only if some other causal factor is present. Thus, the causal effect of smoking on peptic ulcer seems to apply only (or mainly) in individuals who are *Helicobacter pylori* positive (Office of the US Surgeon General 2004).

Second, two way causal influences may operate. There are many examples of children's effects on their parents as well as of parental effects on their children (Bell & Harper 1977). This applies for example, to negative or coercive parenting which is influenced by the behaviour of the children (see O'Connor *et al.* 1998; Anderson *et al.* 1986), but which nevertheless contributes to the causation of psychopathology in the children. See also Section 5.1.

Taken together, these considerations mean that the identification of causes involves a series of substantial challenges. The remainder of the report considers how these may best be met.

Two further points need to be made clear. First, our task has been to determine how to identify individual features with a true causal effect on some disease or disorder. With multifactorial disorders, as we have indicated, it will be usual for multiple individual causes to be involved. Putting all the causal elements together in a total model is an important challenge, but it has to be preceded by identification of the individual causal elements. Our focus has been strictly on this preceding identification and not on how they combine together. Second, our focus has been on the average causal effect in the population and not on the effect as it

applies to individuals. Accordingly, we do not discuss single case research designs.

2.2 Are environmental influences on human disease likely to be important?

The starting point for the setting up of the working party was the apparently high frequency of unsubstantiated claims that some environmental cause for disease had been identified. Before proceeding to discuss how these may be investigated, we need to ask whether there is any reason to suppose that such causes will prove to be sufficiently important to warrant research into their identification? The question is, perhaps, especially necessary to pose in light of the huge advances taking place in genetics. How much space will be left for environmental contributions to causal processes?

Actually, the genetic evidence with respect to multifactorial diseases and disorders clearly indicates that genetic factors never account for all the population variance, and usually non-genetic factors account for some quarter to a half of the variance (Plomin, Owen & McGuffin 1994). Moreover, the effects of genes and environment are not necessarily additive (Rutter 2006). Thus, the inherited metabolic disorder of phenylketonuria is entirely genetic in the sense that it is wholly determined by a particular genetic mutation, but it is effectively wholly environmental in the sense that a dietary intervention removes almost all of the adverse consequences although, of course, it does not change the genetic mutation. Nearly all diseases involve a combination of genetic and environmental influences. Sometimes they are additive and sometimes synergistic. The practical consequence is that the environmental causes may be modified but, in the present state of knowledge, there is less scope for altering the genetic influence.

It has not proved easy to nail down firmly just which environmental features have effects on which outcomes by which mechanisms, but that means that an increased emphasis is needed on high quality studies that could identify causes. That is also the need in relation to the growing evidence on gene-environment interactions (Rutter 2006). The elucidation of genetic causal pathways will be aided by the investigation of gene-environment interactions, but that in turn requires clear identification of environmentally mediated causal effects of specific measured environmental features. In addition, the increasing evidence on the role of environmental influences on gene expression means that the impact of genes may be shaped by environmental forces. This, too, argues for the need to identify environmental causes of disease or disorder.

3 Types of designs used to identify causes

Five broad groups of designs may be used to test for causal effects (Shadish, Cook & Campbell 2002).

3.1 Experiments

First, there are experiments in which some intervention is deliberately introduced in order to observe its effects. For a variety of reasons, these designs give rise to the strongest inference of causation. Their strength lies in the control of the intervention that they provide, plus control of the assessment of effects. Their use is central in laboratory-based science, and that is one reason why basic science plays such a vital role in the understanding of causes. Nevertheless, they have a very limited place in the study of human disease because there are so few circumstances in which it is both feasible and ethical to deliberately give someone the agent postulated to cause disease. There is more scope for experiments in the study of protective interventions. The study of the impact of administration of putative causal agents (such as toxins) is most often accomplished through animal models, and we discuss these separately below.

3.2 Randomised controlled trials (RCTs)

Second, there are RCTs. These were pioneered by Fisher (1925), and first used in agriculture. Since World War II, starting with the trial of streptomycin for the treatment of tuberculosis (MRC 1948 & 1949), they have been used increasingly for medical interventions and they have come to be accepted as the 'gold standard' for determining whether a treatment truly 'works' effectively (Collins & MacMahon 2007; MacMahon & Collins 2001). As Cartwright (2007) has expressed it, they provide 'clinching' evidence of a causal effect within the sample studied.

As with other experiments, the strength lies in the rigorous control of the interventions, with the crucial additional design feature of random assignment to the experimental interventions or to some control condition. The importance of this design element is that it eliminates the serious confound of choice. This is important because of the extensive evidence that people who choose to have a particular treatment, or who have the opportunity to receive it, tend to differ systematically from those who do not make that choice, or who do not have the opportunity. Expressed another way, randomisation ensures that confounding variables are likely to have a zero correlation with the treatment condition. Moreover, the zero correlation will include unmeasured confounders as well as ones that are known to be relevant and which are measured appropriately.

A further key design element is that both the researchers and the recipient of the intervention are kept 'blind' to whether they are receiving the experimental or control intervention. When the intervention comprises a product for which a placebo can be made, the maintenance of a 'double blind' situation is usually possible, in which both the subject and the researcher are unaware of allocation. In sharp contrast, that may not be possible with many public health interventions. It has been argued that non-blind RCTs may be influenced by patient preference (McPherson, Britton & Wennberg 1997) and patient compliance (Simpson *et al.* 2006). That does not mean that RCTs should not be undertaken, but it does call for care in considering patient actions that could create bias.

RCTs are not entirely free of problems (as we discuss in Sections 5.1 and 5.6). For the purposes of this report, however, the main conclusion is RCTs have limited utility because so few hypothesised causal influences are practically open to ethical manipulation. They provide information on the effects of some change in treatment but this may or may not have played a role in the causal processes leading to the development of the disease being studied.

3.3 Regression discontinuity designs

Third, there are regression discontinuity (RD) designs. These were introduced by Thistlewaite and Campbell (1960) nearly half a century ago and were first applied in medicine half a dozen years later (Finkelstein, Levin & Robbins 1966 a and b). The key defining feature is that allocation for some planned interventions is by the assignment variable, using a strict predetermined cut off, rather than randomisation. In other words, the design capitalises on a major selection bias – provided that it is under strict control. The basic point is that the assignment variable cannot have been caused by the intervention; it does not matter whether or not it is related to the outcome. It does, however, require that all participants belong to one population prior to assignment. Thus, for example, the population might be all patients attending a hypertension clinic. The assignment for use of a drug to lower the blood pressure would be some preset cut-off on a standardised measurement of blood pressure. An intervention effect is shown by a difference between regression lines (i.e. slopes on a graph plotting the variable under investigation against the outcome for the various subjects) for the groups above and below the cut-off, rather than a difference between means as in an RCT.

The RD design provides a useful alternative to RCTs for planned, controlled interventions but it is essential that the intervention is based on the chosen cut-off and not clinical judgement. The statistical analysis also requires accurate specification of the intervention effect (e.g. whether it is linear or curvilinear) and also inclusion of an interactive term when this is relevant. Whilst, at first sight, it is not obvious that RD allows an unbiased estimate of a causal effect, it has been shown mathematically that it does (Rubin 1977; Shadish, Cook & Campbell 2002), and this constitutes its major advantage (Laird & Mosteller 1990).

RD designs, like RCTs are limited in their applicability to putative prior causal influences that are not, and cannot be, controlled because it is either impractical or would be unethical. Like RCTs, therefore, they are mainly of use if an intervention to remove a risk effect is possible as a way of testing for the hypothesised causal effect of the risk variable. Nevertheless, occasionally they are applicable to the study of prior causes that are not treatments - see Rutter, (2007 b), for a discussion in relation to Cahan and Cohen's (1989) use of a fixed date of school entry to study the effects of duration of schooling on cognitive performance, and in relation to the use of a discontinuity in programme funding to assess the effects of Head Start on children's health and school progress (Ludwig & Miller 2007).

3.4 Natural experiments

Fourth, there are natural experiments in which, although the cause cannot be ethically manipulated, particular circumstances obviate the allocation bias of individual choice. A description of natural experiments can be found in Box 2. As such, to the extent that allocation bias can be truly eliminated, they approximate to contrived experiments. Because they have important strengths, we discuss them in greater detail in a separate section. They constitute examples of a broader class of quasi-experiments in which there is no random allocation of interventions.

Box 2 What is a natural experiment?

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. For example, adoption separates biological parentage and social rearing. Similarly, a population-wide famine avoids the possible bias created by factors leading some individuals to be malnourished but not others.

3.5 Non-experimental studies

Fifth, there are non-experimental studies that simply observe the size and direction of associations among variables. Because so many of the claimed causes derive from non-experimental studies, we discuss their strengths and weaknesses in greater detail in a separate section of the report. Here we simply note that most consist of cohort studies, case-control comparisons or ecological designs.

3.5.1 Cohort studies

Cohort studies start with a defined population that is then followed up to investigate disease outcomes. They have five main advantages and two important disadvantages. The advantages are:

1. The sequence and timing of associations are readily determined.
2. There is no need to reply on long-term retrospective recall.
3. If properly planned, there should be a better measurement of putative risk factors than is ordinarily possible in a case-control design.
4. They provide a ready estimation of size of effect.
5. There is a good opportunity to examine a wide range of both expected and unexpected outcomes.

This final advantage, however, carries with it the accompanying risk of 'data dredging'. It is not at all that we should expect the effects to be homogeneous across all subjects. To the contrary, heterogeneity is usual. Studies of the gene-environment interactions illustrate well how this may be investigated in a systematic, pre-planned, hypothesis testing fashion (see Caspi & Moffitt 2006; Rutter 2007; Moffitt, Caspi & Rutter 2005). The problem, rather, lies in the undirected attention to an endless list of possible subgroups – a strategy that causes a huge potential for generating false positives.

The two main disadvantages are:

1. Very large samples are required if the disease outcomes to be examined are uncommon.
2. A long time frame is needed to study most associations with disease.

3.5.2 Case-control studies

Case-control studies differ in terms of comparing possible causal factors in individuals with and without some specified disease. If the data are collected at one point in time (cross-sectional) then it can be most difficult to sort out whether the putative cause preceded its supposed disease effect. However, longitudinal data, where the data are collected over time, can remedy that problem. Case-control studies do not have the two main disadvantages of cohort studies (see above), but they are weaker with respect to the five advantages of cohort studies. There is usually a need to rely on long-term retrospective recall, quantification of size of effects involves more tricky assumptions, and necessarily it has to focus only on some specified outcome. The basic point, however, is that, provided that there is no reason to suppose that the two designs rely on fundamentally different cohorts, they are directly comparable statistically (see Cornfield 1951). When possible, there may be advantages to combining the two approaches by 'nesting' a case-control comparison within a broader epidemiological/longitudinal cohort study.

3.5.3 Ecological designs

Ecological studies constitute another non-experimental method that also has a long history in public health research (Susser 1973). Ecological correlations mean the comparison of aggregated groups, rather than the examination of associations at an individual level. Such correlations played a key role in Snow's study of cholera rates in London districts, and in Goldberger's studies relating diet and economic conditions to pellagra (see Diez Roux, Schwartz & Susser 2002). Researchers have sometimes asked what happens to the population rate of a

disease outcome if the environmental risk for it is removed (see Section 6.4.1 for an example where this approach cast major doubt on a causal hypothesis and Section 6.3.3 for an example in which the change over time supported the causal inference).

Ecological associations may also be directly relevant when the postulated environmental cause concerns broader social circumstances such as living in a socially disorganized area (see March & Susser 2006; Reiss 1995; Sampson, Raudenbush & Earls 1997). This field of eco-epidemiology may be particularly informative in identifying broader area influences on the liability to develop a disease, but its use in this way is still developing.

On the other hand, particular caution with ecological correlations is needed because it cannot be assumed that aggregate correlations imply individual correlations (see Greenland 1992 and Robinson 1950, for a discussion of the ecological fallacy). Thus, there are dangers in assuming that a measure of where someone lives is an adequate measure of their individual social status. That is because many personally disadvantaged individuals do not live in a disadvantaged area. Also, the causes at an individual level may be quite different from the associations at an aggregate level. For example, disorders may be more frequent in areas with a high proportion of ethnic minorities because discrimination and housing policies mean that ethnic minorities tend to live in less healthy environments. In this case the risk stems from the broader environment and not the individual person's ethnicity as such. Attention also needs to be paid to the comparable reverse problem; the atomistic fallacy of assuming that causes at an individual level also account for group differences. Ecological correlations can be very informative but particular care is needed in their use and interpretation.

3.6 Animal models

Animal models have a useful role to play in testing causal relationships in the reciprocal interactions among social, behavioural and genetic contributors to health and disease (Hernandez & Blazer 2006). Their utility derives from the possibility to manipulate single variables, or specific groups of variables, in a highly controlled context. Potentially, they provide the opportunity to establish causality through investigations both to examine the temporal sequence of events and involve the removal, followed by the add back, of hypothesised mediators. Such controlled removal and add back can be achieved at the genetic, protein, physiological, behavioural, or social environment level. They allow, also, for invasive examination of organ tissue and region specific mechanisms at the physiological, cellular, and molecular levels. A further advantage of animals with short reproductive cycles and life spans is that developmental and lifespan studies of risk and protective effects are possible in a way that is impractical with humans. Genetic manipulation and breeding experiments facilitate the elucidation of genetic effects, which may be crucial (in terms of gene-environment interactions) for the study of environmental causes of disease.

The essential goal of animal studies is the elucidation of the physiology involved in causal pathways. The findings can, thereby, help in indicating plausible biological mechanisms that might apply in humans. Of course, because of interspecies differences, it is never safe to assume that the findings can be generalised to humans but, equally, it is quite wrong to suppose that all findings are species specific. Rather, the need is to test for generality and specificity going across species.

It is sometimes assumed that, because cognitive and language thought processes are so much more complex in humans than in most other animal species, animal models provide a poor approach for mental disorders.

However, that is to misunderstand what they can, and cannot, do. Of course, it would be difficult to develop a model of, say, autism or schizophrenia in mice or rats, but there may be closer parallels with part functions (such as repetitive behaviour or lack of social engagement) and if it can be shown that biological mechanisms are not contingent upon the availability of cognitive processes found only in humans, it certainly should suggest that simpler mediating mechanisms need to be considered.

Nevertheless, there are limitations. It is important not to interpret an animal's behaviour in human terms without measuring different facets of the behaviour in order to demonstrate which behaviour system is mediating effects. Hernandez and Blazer (2006) used the water maze as an example. This may be problematic. Mice do not ordinarily swim, and navigating a circular pool to find a submerged platform is not something that happens in ordinary circumstances.

A successful performance may reflect spatial learning but also it might reflect other features within the mouse's repertoire.

The value of animal studies in the identification of biological mediating mechanisms involved in the causation of human disease has been discussed in some detail in the Nuffield Council on Bioethics (2005) report on the ethics of research involving animals, and in the Weatherall (2006) report (sponsored by the Academy of Medical Sciences, MRC, Royal Society and Wellcome Trust) on the use of non-human primates in research. For example, a rodent model of rheumatoid arthritis allowed the causal role of TNF (tumour necrosis factor) to be investigated, leading on to testing whether antibodies against TNF could be used therapeutically. Animal models were crucial in studying BSE (bovine spongiform encephalopathy) and its transmission through blood; a chimpanzee model allowed the isolation and characterization of the hepatitis C virus; and monkey models were vital in the study of the spread of polio. Other examples are noted in later chapters.

4 Non-experimental research in medicine

Observations made in the clinic, laboratory, or wider community lie at the heart of the scientific method in biomedicine. Sometimes the observations are serendipitous so that the importance lies in the scientists seeing the significance of something neither planned nor expected. Thus, this applies to Fleming's realisation in 1928 that the spoiling of his culture of staphylococcal bacteria by contaminating mould (which he called penicillin) might carry an important message (see Le Fanu 1999). Curiously, he did not follow through with the needed research to explore the therapeutic potential, but a decade later Florey and Chain did so with experimental research to identify the mechanism involved (a discovery that led to the Nobel Prize in 1945). In other cases the observation constitutes an element in a search for a cause, as with Marshall and Warren's discovery of the pathogenic importance of *Helicobacter pylori* in the causation of peptic ulcers (see Le Fanu 1999) that also led to a Nobel prize. In both of these cases, the initial observation needed to lead on to a range of other research approaches in order to test the causal effect.

The same applies to the observation of the many-fold increase in the risk of vaginal clear cell adenocarcinoma in the daughters of women who used diethylstilbestrol (Harbst *et al.* 1971), the large increase in cardiac valve abnormalities in patients taking fenfluramine and related appetite suppressant drugs (Khan *et al.* 1998) and the even larger increase in the risk of Stevens-Johnson syndrome with anti-epileptic drugs (Rzany *et al.* 1999). In each of these examples, the outcome was rare in unexposed individuals, whereas the excess risk was large in exposed individuals. Vandenbrouke (2004) argued that non-experimental observations might be particularly useful in the identification of unexpected and unpredicted adverse effects of some experience – such as the association between asbestos and mesothelioma, or that between intrauterine radiation and leukaemia in adulthood.

Non-experimental observations can also be useful in identifying powerful treatment effects. This was the case with the finding of the beneficial effects of oral rehydration in treating childhood diarrhoea (Rahaman *et al.* 1979). In this instance, the efficacy was later confirmed in an RCT. In other cases, the effects have appeared so marked that RCTs seemed either unnecessary or impractical. This would apply, for example, to the use of (Glasziou *et al.* 2007):

- Insulin in treating diabetic ketoacidosis.
- Thyroxine in treating symptomatic myxoedema.
- Cortisone acetate in treating Addison's disease.
- Sulphonamides in treating puerperal sepsis.
- Vitamin B12 in treating pernicious anaemia.
- Chloroform in general anaesthesia.
- Defibrillation in treating ventricular fibrillation.
- 'Mother's kiss' to dislodge a nasal foreign body in children.
- Laser beam therapy to treat port wine stains.

These examples have two main characteristics. First, the interventions all have a profound therapeutic effect in circumstances where improvement would otherwise have been expected to be minimal (i.e. a high signal to noise ratio); second, the interventions derived out of sound biological principles.

Non-experimental research has made valuable contributions to the implementation of health technologies. For instance, RCTs have had limited success when evaluating diagnostics, despite valiant attempts, few of which showed much advantage. Knee magnetic resonance imaging (MRI) has become widely accepted as the optimum investigation before diagnostic arthroscopy as a result of non-experimental research (NICE 2007; MacKenzie *et al.* 1996).

As noted, non-experimental observations have played a crucial role across the whole of medicine. In this report, however, we focus

only on such methods as applied to the identification of causes of disease as they concern circumstances in which RCTs are either impractical or unethical. In Chapter 6 we consider a wide range of different examples in greater detail. Here in Boxes 3 and 4 we simply note two examples in which non-experimental methods identified causal components of disease, and in which such identification led to actions with respect to policy and clinical practice, which contributed significantly to improvements in human health. Rather than review the value of non-experimental research in medicine more generally at this point, we simply conclude that it constitutes a key method in the identification of the causes of disease. In Chapter 6, we provide more detailed examples that illustrate both instances in which non-experimental research has led to convincing identification of causes and instances in which this has not been the case.

Box 3 Smoking and lung cancer

Perhaps the most celebrated example of the success of non-experimental research is the discovery of the link between smoking and lung cancer, which was largely based upon the ground-breaking work of Doll and Hill (1950 & 1954). Initially the finding met with disbelief and inaction, but the magnitude, consistency, dose-response relationship and biological plausibility of the association earned it credence. The details are considered further in Chapter 6. Within a decade, confidence in the link was such that the Royal College of Physicians and US Surgeon General published separate reports that identified smoking as a likely cause of lung cancer (Royal College of Physicians 1962; Office of the US Surgeon General 1964).

Since the launch of these reports over 130,000 papers have added breadth and depth to their findings (Royal College of Physicians 2004). Smoking has been in general decline in much of the developed world and it has been estimated that widespread cessation of smoking in the UK since 1950 has approximately halved the mortality from lung cancer that would have been expected if former smokers had continued to smoke (Peto *et al.* 2000). The negotiation of the WHO Framework Convention on Tobacco Control offers further clarification of the contribution rigorous non-experimental research can make to policy and practice (WHO 2003).

Box 4 Cardiovascular disease and high blood pressure

The link between cardiovascular disease and high blood pressure is another success story in which non-experimental findings were key. In the 1960s non-experimental research, such as the widely quoted Framingham study, revealed high blood pressure as a risk factor for heart disease and stroke (Epstein 1996; further information is available from: http://www.framingham.com/heart/).

Many subsequent studies, both non-experimental and experimental, have confirmed the link, which increases the risk to the individual of various cardiovascular consequences two or three times (Padwal *et al.* 2007). Moreover, it has been estimated that the government policies to reduce blood pressure could save 15 million disability adjusted life years (DALYs) per year world-wide (Murray *et al.* 2003) – DALYs being a way of combining the years of healthy life lost through disability and premature mortality. Clearly, understanding of the link between high blood pressure and cardiovascular disease, to which non-experimental research contributed much, has had an important impact on human health.

5 Identification of the causes of disease

5.1 Non-causal explanations of an observed association

Before turning to the ways in which causal inference can be tested, we need to consider what non-causal alternatives have to be examined. The first possibility is that the association simply reflects chance. The choice of a particular significance level as a cut-off (typically a five percent level) means that there is only a low likelihood that the result has arisen by chance. It follows, however, that there is still a possibility of a chance association. The possibility is low if the significance level is very high, but replication is essential if there is a likelihood of systematic bias. It is only when the same association is found repeatedly in different populations and different circumstances that there can be much confidence that it is not a systematic error.

As Rosenbaum (2007) has noted, replication strengthens the evidence only if it removes some weakness in previous studies. This entails varying the evidence rather than just repeating it with the same set of limitations. Much earlier Lykken (1968) made the same point in his discussion of the need for 'constructive replication'.

Publication bias is also an appreciable problem with respect to replication; that is to say, journals are much more likely to publish a positive finding than a negative one. It is quite a common occurrence to find that there is an accumulation of unpublished negative findings that would change conclusions if they had been known.

The second possibility is that the association represents selection (allocation) bias. That is, the association reflects the origins of a risk factor and not its effects. For example, poverty is associated with a wide range of adverse health outcomes (see Section 6.3.2). But, to what extent is that because poverty as such causes ill health and to what extent is it rather that poverty serves as an antecedent of other environmental causes of disease such as smoking?

In other words, is there an indirect distal causal effect because poverty predisposes to some more proximal causal mechanism that actually leads to the disease? Exposure to environmental hazards does not occur randomly. It is influenced by people's selection and shaping of environments, as well as by society's allocation of resources such as housing or employment. Of all the artefactual causes of an association, this is probably the most important. Indeed, it was this consideration that played the main role in the development of RCTs. It was appreciated that people who chose to have a particular treatment or preventive measure would often not be the same as those who chose not to have it. RCTs obviate this problem by ensuring that whether or not individuals have the intervention is determined purely by chance, rather than by choice.

One possible limitation of RCTs is that the people willing to participate may not be the same as those for whom the intervention is being considered. Insofar as that is the case, the price of the experimental control may be loss of their validity with respect to the people to whom the findings need to apply (lack of ecological validity) (Heckman & Smith 1995). It should also be appreciated just because it can be demonstrated that manipulation of some variable influences a particular outcome does not necessarily mean that that variable was involved in the prior causation of the outcome. RCTs are powerful in testing whether an intervention has an effect but, unless specifically designed to do so, they will not necessarily be informative on the mediating mechanism.

The third possibility is that of 'reverse causation', in which, for example, a disease or disorder causes a change in a behaviour that is purported to serve as a cause. The best

known examples of reverse causation concern the many claims that socialisation experiences cause some form of mental disorder. A key paper in 1968 (Bell 1968) pointed out that, in many instances, it was just as likely that children's behaviour was influencing parental behaviour (or the behaviour of teachers) as it was that the rearing environment had caused the child's behaviour. There have been many studies since 1968 demonstrating the reality of this two way operation of causal influences.

Similar concerns have been expressed about the association between schizophrenia and low socioeconomic status (SES). Does the low SES predispose to mental disorder or does mental disorder lead to a drift downward in SES (Miech *et al.* 1999)? Similar concerns arise with respect to SES and somatic diseases (Adler & Rehkopf, in press).

Genetic mediation constitutes a further important possibility (Plomin & Bergeman 1991). This arises as a consequence of gene-environment correlations of one kind or another. A person's behaviour is, in part, genetically influenced, and their behaviour serves to shape and select their environments. As a consequence, many studies have shown that part of the mediation of risk effects from some adverse environment is genetically, rather than environmentally, mediated (see Section 5.5.1).

Yet another possibility when dealing with complex causes is that the association is real and that it does also involve a causative influence, but that the risk element has been wrongly identified. Thus, half a century ago the World Health Organization made strong claims that daycare constituted a very serious cause of mental disorders. The basis for that claim lay in research summarised by Bowlby (1951) that institutional care constituted a significant risk and also that seriously disrupted family life was associated with an increased liability to develop mental disorder. The error lay in supposing that separation was the key risk element and that the brief separations involved in daycare are equivalent to what happened with institutional care (see Rutter 1971). Particularly when dealing with broadly defined risk (or protective) features, there is always a risk that the wrong aspect of the experience is being picked out as responsible for the outcome being studied.

A sixth possibility is that, although the association does reflect a causal influence, this causal effect is contingent upon some particular context. For example, maltreatment in childhood is associated with an increased risk of both antisocial behaviour and depressive disorders in adolescence/early adult life, but the risk effect is largely confined to those with particular genetic variants (Caspi *et al.* 2002 & 2003). Similarly, heavy early use of cannabis is associated with an increased liability to schizophrenia but this appears to be contingent on a particular variant of the *COMT* gene (Caspi *et al.* 2005). Genetic variations also have a substantial moderating role in the risks of cigarette smoking in relation to lung cancer (Zhou *et al.* 2003).

Social context may be similarly important. Thus, care by individuals other than parents seems to be a protective factor for young children living in adverse circumstances but this does not apply to those living in more advantageous circumstances (Borge *et al.* 2004; Geoffroy *et al.* 2007). Similarly, Jaffee *et al.* (2002) found that substantial involvement of fathers in their children's upbringing was psychologically protective for most children, but it provided an increase in risk if the fathers were seriously antisocial.

5.2 Making a causal inference

The first systematic analysis of a causal relationship was provided by the philosopher John Stuart Mill (1843) who argued that three fundamental conditions had to be met. They were:

1. The cause had to precede the effect.
2. The cause had to be statistically associated with the effect.
3. There had to be no plausible alternative explanation for the effect other than the cause.

Bradford Hill (1965), in a now classic paper, discussed seven guidelines that could help in deciding when a causal inference might be warranted. Actually he gave nine but for present purposes we have combined plausibility, coherence, and analogy with known causal associations. Only temporality, the need for cause to precede effect, was essential. The other six were focused on ruling out non-causal explanations. First, the stronger the association, the less likely it was either coincidental or due to confounders. Like many subsequent commentators, Bradford Hill recognised that the causal influences of public health importance might still have only a small effect. His argument, however, was that, when the effect was small, it was much more difficult to rule out confounding. That remains the case.

Second, a true causal effect was more likely when the statistical association was consistent across samples, across different methods of measurement, and across different environmental circumstances. That remains a useful guideline but caution is required because a true causal effect may require complementary component causes.

Third, it was suggested that a biological (including psychological) or dose-response gradient helped in making a causal inference. The reason is that, if such a gradient represented a non-causal effect with respect to the postulated causal influence, it would need to be due to a confounder that showed the same dose-response gradient.

Ordinarily, that was likely to mean a true causal effect from the confounder, but there are exceptions. Thus, Rothman and Greenland (2002) gave the example of the gradient found with respect to birth rank and the incidence of Down syndrome. There is no causal effect of birth rank as such, but there is a causal effect from a higher maternal age, which will be associated with birth rank. A biological gradient will also not apply if there is a threshold effect such that there is a cut-off above or below which the cause had little or no effect on the outcome in question. Rothman and Greenland used the example of diethylstilbestrol and adenocarcinoma of the vagina as an illustration. In this context, it is not well understood why there is a threshold effect but there seems to be one.

Fourth, attention should be paid to plausibility or coherence in terms of well established scientific knowledge on both the postulated risk experience and on the causal processes involved in the disease being studied. Insofar as there is plausibility, it provides some support for the causal inference. It is limited, however, by the fact that most creative scientists are skilled in putting forward suggestions on how a cause might operate. Bradford Hill was not referring to such an hypothesised mediating possibility. Rather, the suggestion was that there should be scepticism about the causal inference when it seemed to run counter to existing knowledge or when it seemed to have no plausible mediating mechanism. Of course, that cannot completely rule out a true causal influence because existing knowledge may be wrong or incomplete.

Fifth, the causal inference is more likely to be correct if experimental or quasi-experimental evidence is available to test at least one crucial part of the hypothesised effect; Chapter 6 gives several examples in which this has been the case. It was appreciated that with many causes this was just not practical but, insofar as it was, it was very supportive of the causal inference. Expressed more broadly, the implication was that the inference was more likely to be correct if it was supported by several different research strategies. The sixth element to the guidelines was that the effects should be *specific*. Rothman and Greenland (2002)

picked this out as the one guideline that was invalid – on the grounds that many genetic and environmental causes had pleiotropic (meaning multiple varied) effects. The example of smoking would seem to support crossing this off the list of guidelines because of the good evidence that smoking has such pervasive negative effects on health (see Office of the US Surgeon General 2004). We agree that this is the weakest of the guidelines but its weakness lies mainly in our understanding of mediating mechanisms. Once these are known, there may be more specificity than at first apparent. Thus smoking does not involve just one environmental hazard; its effects may derive from carcinogens, from nicotine, from carbon monoxide or from physical irritation - to mention just some of the possibilities.

Similarly, stress has widespread effects on psychopathology but there may be some specificity through the causal route involving immune mechanisms or neuroendocrine effects. The valid guideline is that the causal inference is much strengthened if the mediating mechanism can be identified and tested. The inference of causation does not require such identification but it is more secure if that is possible.

We conclude that, with minor adjustments, the Bradford Hill list remains an excellent set of guidelines (Phillips & Goodman 2004). He was explicit that they should not be treated as rules, or given a score, but rather they need to be seen as a way of thinking about how to proceed from a statistical correlation to a causal inference. We agree. Following John Stuart Mill, Schwartz and Susser (2006) emphasised the importance of ruling out alternative sources of differences between exposed and non-exposed groups. They pointed to the value of 'natural experiments' in contributing to this ruling out process and we, too, discuss their role. Such research strategies may also help 'ruling in' if they can identify 'footprints' by which the causal effect might be recognised if it were real.

5.3 Counterfactual reasoning

All causal reasoning requires an implicit comparison of what *actually happened* when individuals experienced the supposed causal influence with what *would have happened* if simultaneously they had not had that experience. Clearly, that observation can never be made, even in an experiment. The control provided by an experiment can show the effect of the intervention on the individuals who received it – by comparing pre-test and post-test measures. Because experiments are by design longitudinal, *within individual change* provides a convincing demonstration of the effects of the intervention. Nevertheless, there is still the need to determine whether such change might have taken place anyway even if the intervention had not been experienced. Hence, all experiments (whether in the laboratory or through RCTs) include some form of control condition.

The randomisation procedure in RCTs ensures that the two groups (experimental and control) are comparable. It does not ensure that there will be no relevant differences between the two; indeed, by chance, these are bound to occur occasionally. Accordingly, statistical techniques will need to be used to take those into account and to rule out the possibility that these chance differences brought about an artefactual difference between the outcomes in the two groups.

What RCTs guarantee, however, is that any confounding influences will be equally likely to arise in both groups. Most especially, this equality applies as much to unknown and unmeasured confounders as to known measured confounders. No amount of matching in non-experimental designs can achieve that. It is because of that fact that RCTs have come to be viewed as providing the 'gold standard' for causal inference. We consider in Section 5.6 the extent to which this is justified but first, we discuss the meaning of counterfactual reasoning.

A counterfactual is something that could have happened, but not simultaneously with the exposure to the supposed causal influence – in this case what would have happened if the exposure had not occurred. Note that counterfactual reasoning is necessarily judgmental (even in experiments). Nevertheless, it has come to be widely employed in thinking about causes (see Hernán *et al.* 2004; Mackie 1974; Maldonado & Greenland 2002; Rubin 2004). Moreover it has also given rise to statistical models (see Rubin 1986). Most researchers have come to view counterfactual reasoning as crucial in drawing causal inferences (although reservations have been expressed – see Dawid 2000; Pearl 2000). Regardless of these concerns, it is essential to consider what is required to determine if some effect represents a true causal influence.

5.4 Dealing with errors and confounders

Non-random (or systematic) errors are different from random error in that they give rise to bias, see Box 5 for a description of random and systematic error. Experiments (including RCTs) differ from observations, not only in introducing an imposed intervention that is compared with no intervention (or a different intervention) in terms of its effects on some pre-specified outcome, but also in the careful steps taken to control for the effects of all factors other than the intervention being tested.

The basic principles were laid out by Fisher (1925) nearly a century ago and still apply today. In laboratory studies, or animal investigations, it is generally easier to do this than it is in experiments or observations with people. That is why randomisation of the intervention tends not to be done in basic science experiments. Nevertheless, some of the difficulties in translating animal models to man may be due to lack of attention to biases, as well as interspecies differences. Thus, interventions in animals that have been effective in reducing the effects of strokes or myocardial infarction were found to be ineffective in humans.

Box 5 Random and systematic error

All forms of data gathering, either within an observational or experimental framework, involves some form of sampling. That is, some subset of a population of interest is used to represent all possible observations or an infinite number of experiments. All types of study yield findings that might have arisen by chance, thereby producing random errors. The available safeguards include adequate sample sizes, replication, the testing of only pre-specified hypotheses, and the application of considerable caution when approaching findings from some subset of data. These considerations apply equally to observational and experimental work and erroneous claims of causality arising from random errors can derive from any form of research. Moreover, the history of research clearly indicates that many of the mistakes in the past derive from small studies that threw up false positive findings.

5.4.1 Major sources of bias in non-experimental studies

Collins and MacMahon (2007; MacMahon & Collins 2001) have provided a detailed critique of the major sources of bias in non-experimental studies. Three substantial problems stand out.

First, non-experimental studies are especially prone to bias when there is a marked selection effect (leading to *allocation bias*); that is, the individuals who opt for some intervention differ in a major way from those who do not, and where this difference is associated with the outcome being investigated. The use of HRT in relation to a supposed protective effect on coronary artery disease constitutes the most striking example of this kind (Beral *et al.* 2002; Hsia *et al.* 2006; Manson *et al.* 2003; Pahor *et al.* 2000) – see paragraph 6.4.2.

Second, there is a similar likelihood of bias when a treatment intended to be protective is differentially prescribed for high risk patients – *indication bias*. The use of calcium antagonists to reduce the risk of heart attacks is the striking example of this problem (Blood Pressure Lowering Treatment Trialists' Collaboration 2000; Psaty *et al.* 1995) – see Section 6.4.3.

Third, bias is probable if the individuals receiving any treatment will tend to be seen by professionals more frequently than will others – *ascertainment bias*. This might have applied to the finding of an increased risk of breast cancer in women taking hormonal contraceptives (Collaborative Group on Hormonal Factors in Breast Cancer 1997), and to the increased risk of congenital malformations in the children of women taking the antifungal drug itraconazole (Bar-Oz *et al.* 1999). It will be appreciated that all three biases have been shown with respect to treatments. Possibly, they may be less likely to apply in other circumstances, but it would be foolhardy to count on that.

The lesson should be that particular attention should be paid in all non-experimental studies to the possibility of allocation bias, indication bias, and ascertainment bias.

5.4.2 Confounders

All studies seeking to identify causes need to pay attention to the possibility of effects of confounding variables, which are described in Box 6. The need is somewhat less pressing in RCTs because the process of randomisation ensures that confounders should be similarly distributed in the experimental and comparisons groups. Nevertheless, by chance, confounders may be more frequent in one group than another, and such random error will need to be taken into account in analyses. The problems are vastly greater in non-experimental studies for two main reasons. First, because there is no randomisation, it is highly likely that confounders will not be similarly present in the two groups being studied. Second, even more seriously, any attempt at controlling for confounders

Box 6 What is a confounder?

A confounder is any feature other than the hypothesised disease-causing influence that might artefactually give rise to the supposedly causal associations. Such features have to both differentiate the groups and involve an association with the outcome of interest. Sometimes, the confounder represents a causal influence on the likelihood of exposure to the supposed disease-causing influence. This, as we discuss below, was a crucial biasing factor in the study of HRT for menopausal women. Sometimes, however, the confounder may not affect selection into the exposure but rather it may reflect biases in the ascertainment or measurement of the outcome that differ across treatment groups. Alternatively, it may be just some 'third variable', that happens to be correlated with the supposed causal influence and which creates an artefactual association because it influences the outcome. This might apply, for example, to age or sex or genetic effects.

A key challenge in any non-experimental study is how to deal effectively with the possibility that the association might be due to one or more confounders. Note that there should not be exclusive reliance on the statistical significance of individual confounders or even groups of confounders (because that is so strongly affected by numbers). Often it may be preferable to pay more attention to standardised differences between groups before and after taking confounders into account, because there may be possible important confounding effects of even infrequent features. For a very basic simplified account of confounders and some straightforward ways to deal with these, see Mamdani *et al.* 2005; Normand *et al.* 2005; Pickles, in press; Rochon *et al.* 2005.

will necessarily be reliant on those that can be identified and how well they are measured.

Latent variable structure equation modelling (SEM) provides one framework that can be very helpful in the context of imperfect measurement, separating measurement of the underlying latent construct of interest from the variance due to the unique and random error components of whatever measure is being employed (see e.g. Bollen 1989; Pearl 2000; Reichart & Gollob 1986). Expressed simply, what are required are multiple measures of the same latent construct. By using the intercorrelations among them, it is possible to infer an error free estimate of the construct in question. Of course, there is the remaining important question of the validity and appropriate labelling of the construct (e.g. does it reflect the methods being used or the trait in question?). Also, whilst latent variable methods can do much to deal with errors in measurement, on their own they cannot take account of the effects of unmeasured and unconceptualised confounders.

While latent variable methods can do much to deal with errors in measurement, their use does not by itself overcome the problem of unmeasured and unconceptualised confounders. Nevertheless, with suitable data, structural equation models with latent variables can be constructed to provide a more rigorous basis for causal inference. With repeated data measurements, latent variables representing subject specific effects can be conceptualised as the net effect of residual confounders (see Paragraph 11 in Appendix I). Such effects can draw on the association between change in exposure to the putative causal factor and change in the outcome to estimate causal effects. This approach can also be conceived of as an example of the much broader *instrumental variable* approach (see Section 5.5.2) in which specific assumptions about one part of the causal mechanism imply restrictions that allow residual confounders to be represented as a latent variable elsewhere in the model. Latent variable approaches can also be used for the representation of subject specific effects that can be conceptualised as the net effect of residual confounders (see Paragraph 17 in Appendix I). Such effects can draw on the association between change in exposure to the putative causal factor and change in the outcome to estimate causal effects. In the latent variable framework, these are equivalent to instrumental variable approaches (see Section 5.5.2).

Natural experiments (see Section 5.5), each of which has its own statistical requirements, provide a further measure of testing causal inferences using non-experimental methods. This means that the researcher does not manipulate the putative causes but, rather, uses naturally occurring situations to provide variations that are outside the control of the individual.

Two rather different issues arise with respect to adjusting for confounders. First, the adequacy of adjustment will be very dependent on how well the confounder has been measured. Sometimes there is reliance on some general index that is available and is readily measured (such as social class to cover all possible lifestyle differences). The strength of effect (such as reflected in an odds ratio) may be reported as having been adjusted. If use of such a weakly measured confounder markedly reduces the effect of the putative cause, but with some remaining significant effect, there is a strong likelihood that more adequate measurement would eliminate the causal effect. If, on the other hand, the adjustment makes little difference, confounding is less likely. The second issue is whether there are other likely confounders that have not been taken into account at all. As we note in Section 5.4.4, these need to be considered on the basis of what is known about the disease being investigated. The general statistical strategies for dealing with confounding variables are discussed in Appendix I of this report and are discussed in greater detail in Rosenbaum 2002; Rothman & Greenland 1998; Shadish, Cook & Campbell 2002; Susser *et al.* 2006. Here we mainly focus on four issues:

mixed approaches making use of design features; the value of modelling possible causal pathways, the use of propensity scores and sensitivity analysis.

5.4.3 Mixed approaches

Robins (2001) has argued for the crucial importance of both study design and background knowledge about subject matter that provides information about how confounders could lead to a misleading inference. Two examples illustrate the utility of mixed approaches based on design features that are shaped by background knowledge.

Case, Lubotsky and Paxson (2002) tackled the question of whether low household income had an adverse causal effect on children's health. Numerous studies had shown that there are consistent statistical associations and that these become more pronounced as children age. Their main data base was derived from four large scale US surveys. The initial analyses were, as expected, in line with previous non-experimental evidence. However, there were several important likely sources of bias.

Were the effects due to the children's poorer health at birth (due to factors such as poorer prenatal care, maternal smoking, etc.)? Alternatively, were they a function of parental health (influenced by genetic factors that had consequences for the child, or by lower quality care from such parents)? Were the effects due to parental income being a proxy for the genetic tie between parent and child? These questions, and a range of other possibilities, were tested for systematically in a variety of statistical regression based models. In addition, the last mentioned (genetic mediation) possibility was examined by comparing the effects for children living with both birth parents and those living with two non-birth parents – with no difference found between the two.

Rather than rely on a general controlling for confounders, substantially greater leverage was obtained by undertaking hypothesis driven analyses focusing on alternative causal pathways other than those stemming from the effects of family income on child health. In each instance, of course, it was necessary to check that other causes and confounders were equally distributed in the groups to be compared.

The second example concerns Kim-Cohen *et al.*'s (2005) study of the possible environmentally mediated causal effect of maternal depression on children's antisocial behaviour. Their data were based on a longitudinal epidemiological twin study. Other evidence suggested that either genetic or environmental mediations were possible – the former because of the consistent evidence that antisocial behaviour has a substantial (circa 50%) heritability, and the latter because of the evidence that maternal depression affects family functioning and parenting.

The standard across twin, across trait, analysis, making use of the monozygotic-dizygotic difference, could not be used because maternal depression was a feature that constituted a comparable risk factor for both types of twin pair and both twins in any given pair. Accordingly, Kim-Cohen *et al.* used knowledge on possible causal pathways to test competing alternative explanations. Thus, the association could derive from other psychopathology in the mother (because comorbidity is common) and it could reflect assortative mating with antisocial men.

Longitudinal data showed that it was only maternal depression arising after the child's birth that had an effect (making pure genetic mediation unlikely), and they also showed a dose-response relationship with the frequency of the mother's depressive episodes during the child's lifetime. The association between maternal depression and child antisocial behaviour remained after controlling for both maternal comorbidity and psychopathology in the father.

Strikingly, the association was found even in mothers who themselves had no antisocial symptoms. The twin designs also enabled

an analysis of whether the within individual change in the children's antisocial behaviour was genetically or environmentally mediated (rating bias possibilities being dealt with by using both parent and teacher reports separately and together). As with the parental income example, the analysis of possible confounders was greatly helped by focusing on specific mechanisms and by the design advantage of a twin sample.

5.4.4 Statistical modelling based on causal graphs

Robins (2001; Gill & Robins 2001; Robins *et al.* 2000), following Pearl (1995), went further in showing how causal graphs that spelled out the implications of background knowledge could lead to statistical modelling that could go a substantial way in increasing, or decreasing, the likelihood of a causal inference being correct. In both cases, of course, it is essential to consider whether they are a true representation of reality. The illustration of the controversy over whether postmenopausal oestrogens had an effect on uterine endometrial cancer was one of those used. It was biologically plausible that there was a true causal effect but, on the other hand, clinical knowledge indicated that vaginal bleeding was likely to lead to the ascertainment of previously undiagnosed cancer. Both matching in a case-control design and matching according to the presence/absence of vaginal bleeding in the month before diagnosis were shown to lead to biased results.

In other cases, the focus on concrete likely alternative pathways can lead to statistical modelling that can go an important way along the path of testing the causal inference. The key need in all cases is to make explicit the assumptions, their implications, and how they may be tested. The value of these approaches is obviously greatest when there is a great deal of knowledge on biological mechanisms and it is least when little, if anything, is known about confounders. No approaches other than RCTs and RD designs can take adequate account of confounders that are either unknown or unmeasured, unless combined with some form of natural experiment or other design feature that allows causal inference to be tested (see Appendix I, Section 11).

5.4.5 Propensity scores

The third approach involves the use of propensity scores, as advocated by Rosenbaum and Rubin (1993 a & b). Propensity scores reflect the conditional probability of being exposed to the postulated causal agent, given relevant background variables. In many respects it relies on the same analytic techniques as the more familiar regression analyses (Winship & Morgan 1999). However, the rationale is somewhat different in that it seeks to equate groups on risk for exposure to the putative cause (or treatment) being considered, rather than just control for confounders in risk.

If propensity scores are to do the job for which they are designed, the variables included should cover covariates that are found to predict the exposure under investigation (and which might thereby serve to constitute confounders for the outcome). This means that investigators need to consider conceptually what might be happening to lead to differences in exposure.

Note that because RCTs equate groups for all confounding variables, then commonly no matching approach (including propensity scores) should be (or need be) combined with it. Nevertheless, sometimes matching can improve the precision of the treatment effect and when, by chance, the randomised groups may differ with respect to confounders, these should be taken into account in the usual way.

Note, too, that the aim of propensity scores is not to equate on outcome (which should not be taken into account in creating a propensity score), but rather to equate on risk for exposure to the treatment (or hypothesised causal influence) in order to assess its possible causal impact. Propensity scores can then be used to create strata to equate the groups.

One very important advantage of this statistical approach is that it makes it obvious where there are major differences between the groups in exposure to risk. This method, like any other, cannot be expected to work well when there is little overlap between groups in the strata (Shadish, Luellen & Clark 2006). It will usually be desirable to drop strata where there are very few subjects in either the cases or controls to be compared in the quasi-experiment. Propensity scores also form the basis of the weighting used in the marginal structural modelling approach of Robins and colleagues, that can also be applied to longitudinal data. As Robins, Hernán and Brumback (2000) noted, propensity scores may work less well in the case of non-dichotomous exposures, though something similar is possible using g-estimation of structural nested models (SNMs).

Proponents of this approach have been careful to point out that it is no cure-all. In particular, it cannot deal with confounding variables that have not been measured. Moreover, the procedure will be influenced by how the propensity scores are calculated and it is necessary to appreciate that propensity scores have no absolute validity. That is, they will vary according to the particular samples to be compared and the particular variables included. Nevertheless, the one study that compared the efficacy of propensity scores as compared with randomisation was somewhat reassuring (Luellen, Shadish & Clark 2005; Shadish, Luellen & Clark 2006).

In brief, the study involved a two step randomisation into a study of mathematics training or vocabulary training. The first step involved randomisation into either the randomised experiment or the non-randomised experiment. The second step, within the randomised experiment, involved random assignment to mathematics training and vocabulary training. Initially, the non-randomised experimental group included a much higher number of individuals who volunteered for vocabulary training than who volunteered for mathematics training. Not surprisingly, therefore, the initial findings were rather different in the RCT and the quasi-experiment. The use of propensity scores, however, brought them quite close together. It was just one experiment, and it involved psychology students and non-medical outcomes. It cannot necessarily be presumed that the same would apply in other circumstances, but, so far as it goes, it points to the potential utility of the method.

The main use of propensity scores up to now has been to equate cases and controls in quasi-experiments. However, they may also be employed in investigating within individual change over time (see Section 5.5.4). Thus, Sampson *et al.* (2006) used propensity scores to examine whether crime rates varied according to whether individuals were in a marital relationship (see Section 5.5.4). A previous study of theirs had examined this issue with respect to between group differences and found quite a strong marriage effect. The question then was whether it had held up with respect to variations in marital status over time.

The combination of propensity scores and age curve variations in crime rate, with the statistical technique of inverse probability of treatment weighting (IPTW), devised by Robins *et al.* (2000), was used to examine the effects of marital status. It was found that the within individual change findings closely paralleled the earlier between group findings. Because of the strength of the variations, it was important to examine whether the findings held up over both short and long timespans. Analyses showed that they did.

To date, there has been far too little testing of propensity score strategies for any confident conclusion on the extent to which they do control for allocation bias. Some have argued that propensity score methods' advantages over regression techniques are more apparent than real (Avorn 2006). That may be so, but

the advantages of identifying areas of non-overlap remain. The problem of controlling for confounders is much greater when samples are markedly different in their composition (Rubin 1979). The implication is to use groups that are as comparable as possible. It is obvious that any statistical control for confounders in non-experimental studies can only be as thorough as the measures obtained. Solberg *et al.* (2005) in a study of surgery showed the extent to which reliance on crude matching data may create a false sense of confidence.

5.4.6 Sensitivity analyses

A further check is provided by sensitivity analyses (Cornfield *et al.* 1959). In essence, these quantify how strong a confounder would have to be to overturn a causal inference from a case-control comparison. When this was done with respect to smoking and lung cancer, it was found that only a confounder that was nine times as frequent in heavy smokers as non-smokers could undermine the causal inference. Careful consideration of the possibilities indicated that that was extremely implausible; it was much more likely that smoking had a true causal effect.

The final point concerns the need for diverse strategies and diverse samples likely to differ in their patterns of confounders (Susser 1973). Note that the answer does not rely just on replication. If the replications include the same biases, the result will simply be confirmation of a wrong inference (see Rosenbaum 2001). That is where 'natural experiments' are particularly helpful because they pull apart variables that are ordinarily associated.

5.4.7 Can statistical control for measured confounders be sufficient?

Non-experimental research can only take account of confounders that have been both conceptualised and measured. The key question is whether, if this has been thorough and thoughtful, it is possible to assume that bias has not been created by unmeasured confounders? Some leverage on this question is provided by the comparisons with both RCTs and natural experiments. The RCT comparison is informative because, unlike non-experimental methods, it is able to deal with unmeasured confounders. As already discussed, there are very few opportunities to make such a comparison because so few possible environmental causes of disease can be randomised in humans. However, there are a few. In Section 6.3.1 we note the example of the association between HRT and the risk of coronary artery disease as one such instance. A failure to properly account for length of exposure both within, and between, studies also contributed to inconsistency (Prentice, Pettinger & Anderson 2005 and discussion).

As already noted, there are not many circumstances in which the results of non-experimental studies can be compared directly with RCTs, but there are a few (Benson & Hartz 2000; Concato *et al.* 2000; Kunz & Oxman 1998; Pocock & Elbourne 2000). Unfortunately, many of the comparisons are not exact and reviewers have differed in their conclusions on the extent to which the two types of design agree. For the most part, the disagreements concern degree of risk effect, rather than the direction of influence. On the whole, large-scale well conducted non-experimental studies have given rise to findings that are in the same general direction as RCTs. Where they have not, the usual feature has been a small effect and a large selection bias (as noted above).

Comparisons with natural experiments (see Section 5.5) provide another test. Section 6.4.6 provides the example of the claim that an unusually early use of alcohol creates an increased risk of later alcohol dependency or abuse. Several different types of natural experiment showed that the observed associations probably reflect a shared genetic liability rather than environmental mediation of risk. In this instance, as with others discussed in Section 5.5, statistical control for measured confounders did not prove adequate to deal with the known likely allocation bias created by genetic risk.

As with the HRT and coronary heart disease examples, the lesson is that there needs to be considerable caution in non-experimental studies when there are known features likely to create major confounding and when such features can only be partially indexed by measurable variables.

There is one further issue with respect to the statistical control for confounders. That is that it may be unsatisfactory when the groups to be compared show very little overlap with respect to key risk characteristics (see Section 5.4.6), or when the causal effects are crucially dependent on contextual qualities such as gene-environment interactions (see Section 6.2.11). As always, the first lesson is that the statistical approaches need to be guided by a background knowledge of likely mechanisms (see Section 5.4). The second lesson is that causal inferences need to be tested by multiple research designs and not just one (see Section 6.1).

5.5 Natural experiments

Hill (1965) argued that a key need in testing causal inferences was to conceptualise and consider possible alternative explanations for the observed statistical association or correlation. It is never acceptable simply to try to provide supporting evidence that might bolster the causal inference (Shavelson & Towne 2002). As Cochran and Chambers (1965) noted, this requirement is the one most often missing from research into the causes of disease.

Campbell and his colleagues argued for the potential value of quasi-experiments or natural experiments – meaning design elements that provided an approximation to experimental conditions (Campbell & Stanley 1963; Cook & Campbell 1979; Shadish, Cook & Campbell 2002). They noted, amongst other things, that what was needed were designs that pulled apart variables that ordinarily go together.

More than a dozen examples of different forms of natural experiments used in the field of psychopathology have been described and critically evaluated by Rutter *et al.* (2001; Rutter, 2007 b). Here they are noted more briefly. Their importance lies in their power to reduce the risk of bias associated with different types of confounders. Their rationale does not always require longitudinal data but they help considerably. That is because longitudinal data (or data that reflect longitudinal change) are ordinarily required in order to show within individual change; statistical techniques must be used that can differentiate between real change and measurement error (see e.g. Fergusson *et al.* 1996; Sampson & Laub 1996; Zoccolillo *et al.* 1992).

5.5.1 Genetically sensitive designs

Five types of natural experiments focus particularly on the need to differentiate between genetic and environmental mediation of risk effects, although in order to be effective they must also deal with other threats to validity – such as temporal order. This may be done by means of a multivariate twin design that includes both across twin and across trait analyses (one of the 'traits' being the postulated risk factor). For example, numerous studies have shown a clear, and quite substantial, association between an early age of first drinking and the later development of alcoholism (Grant & Dawson 1997).

Multivariate twin analyses, however, showed that there was no evidence that the age of first drinking alcohol had any environmental causal effect on the likelihood of developing alcoholism. Rather, the statistical association reflected genetic mediation. Jaffee *et al.* (2004) used a comparable approach to contrast the effects of both physical punishment and overt maltreatment on the likelihood that the child would develop antisocial behaviour. The findings were striking in showing that the main effect of corporal punishment was genetically mediated (probably reflecting a parental response to the child's behaviour) whereas the effects of

maltreatment were environmentally mediated. Discordant twin pairs can also be used to test for environmental mediation effects. This strategy, like the multivariate analysis approach, showed that there was no environmentally mediated effect of drinking at an unusually early age as an effect on later alcoholism (Kendler & Prescott 2006). On the other hand, the same strategy provided strong evidence that child sexual abuse had a substantial environmentally mediated effect on the risk of later alcoholism and substance use disorders. Discordant twin pairs obviously cannot be used satisfactorily to examine prenatal risk effects but discordant sibling pairs can serve a somewhat similar purpose. Thus, D'Onofrio *et al.* (in press) compared the outcome in pregnancies when the mother smoked and those in which she did not. The findings confirmed the environmentally mediated prenatal effect on birth weight but did not confirm an environmentally mediated prenatal effect on the offspring's antisocial behaviour.

Adoption/fostering designs can also separate possible genetic and environmental mediation effects. Thus, Case *et al.* (2002) compared the effects of low income on the health outcomes of children according to whether or not the rearing was by biological or non-biological parents. The finding that the associations were similar in the two groups pointed to environmental rather than genetic mediations.

Yet another use of genetic designs to test for environmental mediation is provided by the children of twins strategy (D'Onofrio *et al.* 2003; Silberg & Eaves 2004). The rationale is that the offspring of adult monozygotic twins are social cousins but genetic half siblings. The design requires a very large sample and it suffers from the limitation that, ordinarily, there will not be adequate data available on the spouse of each twin. Nevertheless, using this approach, it did seem that harsh forms of physical punishment had an environmentally mediated influence on both disruptive behaviour and drug/alcohol use (Lynch *et al.* 2006). Migration strategies constitute another useful 'natural experiment' for separating genetic and environmental effects. The test is whether, when an ethnically distinctive group moves from a country with relatively low rates of a particular disease outcome to a country with a relatively high rate of some disease outcome, the disease alters in relation to changes in lifestyle, or whether it remain the same in keeping with the fact that the individuals bring the same genes with them. An early study of coronary artery disease in people of Japanese origin living in Japan as compared with California showed that when Japanese people adopted a Californian lifestyle, their rate of coronary artery disease rose to much the same levels as those of Caucasian individuals living in California (Marmot & Syme 1976).

Recently, the same strategy has been used to examine the risk and protective factors involved in the raised rate of schizophrenia spectrum disorders in people of Afro-Caribbean origin living in either the UK or the Netherlands. The findings have shown that the raised rate of schizophrenia spectrum disorders is apparent not only in comparison with individuals of Caucasian origin living in the UK or Netherlands, but also those of the same ethnic origin living in their country of origin. The findings suggest environmental mediation involving some aspect of the adversities associated with migration to the UK and Netherlands (Jones & Fung 2005). A parallel study in the United States has shown a somewhat similar ethnic effect, with African Americans about three times as likely as white people to be diagnosed with schizophrenia (Bresnahan *et al.* 2007).

5.5.2 Other uses of twin and adoption designs

There are some half a dozen other uses of twin and adoption designs for purposes other than separating genetic and environmental mediation. For example, so called 'Mendelian randomisation' was originally introduced in order to deal with the possibility of reverse causation and the approach has received a good deal of publicity in recent years.

The rationale was first outlined by Katan (1986) and has been more fully developed by Davey-Smith and Ebrahim (2003). The ingenious point is that it is possible to use a control provided by a genetic factor that has a strong influence on the independent variable to be examined, but which has no direct association with variation in the dependent variable. This strategy is not concerned with eliminating genetic mediation as a possibility, but rather uses genetic variance as a means of avoiding a confounded or biased association with the disease outcome. What is necessary is that the genetic variant is related to the risk exposure of interest but is not related through any other pathway to the outcome – see Davey-Smith (2006).

A good example is provided by a genetic variant concerned with the oxidisation of alcohol to acetaldehyde, which is strongly associated with alcohol consumption. Japanese people with the variant that renders consumption of alcohol very unpleasant because of marked facial flushing have a substantially lower rate of both alcohol consumption and liver cirrhosis. This genetic instrumental variable can then be used to determine whether the apparent protective effect of alcohol against coronary artery disease holds up using the genetic variant as a control. The point here is that the variant is related to alcohol consumption but is not related to coronary artery disease. The findings have suggested that there is a true (albeit modest) environmentally mediated protective effect of alcohol against coronary artery disease. The strategy has an important utility but this is constrained by the need to have genetic variants with a relatively strong and specific effect on the relevant outcome. When genetic effects are weaker, the design is still applicable but it is likely to require an enormous sample: ordinarily this will mean a meta-analysis – with the usual concerns about comparability across studies. Nevertheless, it has had some successes in providing support for a causal hypothesis (see e.g. Casas *et al.* 2005, in relation to homocysteine levels and stroke, and Brennan *et al.* 2005, in relation to a possibly positive effect of cruciferous vegetables on lung cancer); as well as casting doubt on the effects of fibrinogen levels on coronary artery disease (Keavney *et al.* 2006).

There are several tricky assumptions that are involved in the use of the Mendelian randomisation strategy (Didelez & Sheehan 2005; Meade, Humphries & De Stavola 2006; Nitsch, Molokhia, Smeeth, De Stavola, Whittaker & Leon 2006; Tobin *et al.* 2004). The strategy is innovative and has important strengths but there are uncertainties over the range of circumstances in which it can be used effectively.

Adoption/fostering designs can also be used to separate prenatal from postnatal effects. Thus, Moe (2002) examined the outcome in a sample of babies exposed to drugs or alcohol in pregnancy but who were removed from their mothers' care in early infancy and adopted or fostered. The findings showed a significant effect on cognitive functioning at four and a half years, in comparison with the control group. This effect pointed to a prenatal, rather than a postnatal, adverse effect (the latter being a real possibility with heavy drinking or alcoholic mothers).

When adoption involves a move from a severely depriving environment before adoption to a good rearing environment afterwards, the rapid and radical change in rearing circumstances also provides an opportunity to compare the effects of pre-adoption and post-adoption environments (see below; also Rutter, 2007 b). When nutritional levels vary greatly, the same basic designs may be used to compare the effects of severe subnutrition and severe psychological deprivation in the context of nutritional levels within the normal range (Sonuga-Barke *et al.*, submitted) – with findings that suggested that the main lasting deficits were more affected by psychological than by nutritional deprivation. A rather different type of natural experiment using twins concerns the comparison of twins and singletons to examine the nature of the environmental effects on language delay (Rutter

et al. 2003; Thorpe *et al.* 2003). The possibility of genetic mediation is sidestepped because, although genetic influences will operate within both twin and singleton samples, there is no reason to suppose that those involved with language delay (the outcome being examined) will differ between twins and singletons. Rather, the two main alternatives to be considered were whether the overall delay in language in twins of about three months at three years was mainly due to obstetric/perinatal risks or postnatal differences in parent child interaction. The findings were clear cut in pointing to the latter being responsible for the risk mediation (in a sample with a gestational age of 34 weeks or greater).

The use of a factor external to the liability, but one that influences the risk factor being considered, to the disease or disorder outcome constitutes a strategy similar to Mendelian randomisation, except that the instrumental variable is not a gene. Because puberty is a strongly genetically influenced feature, it is most conveniently considered here, although it does not involve either twins or adoptees. It has been used, for example, to examine the possibility that very early use of alcohol creates an environmentally mediated risk for later alcoholism. An early onset of puberty in girls constitutes the instrumental variable because it is associated with early use of alcohol, although there is no reason to suppose that it has a causal effect on alcoholism that is independent of early drinking. The follow up into adult life in three large-scale general population studies have been consistent in showing that, despite the strong effect on drinking in adolescence, there is no effect on alcoholism in early adult life. The implication once more is that early drinking reflects a shared liability to a broader range of problem behaviours, rather than a causal influence as such (Rutter, 2007 b).

5.5.3 Designs to avoid selection bias

A further group of natural experiments has as their main aim the avoidance of selection bias. RCTs achieve this by means of randomisation and the natural experiments do so by focusing on circumstances in which the experiences apply to all individuals in the group studied without the possibility of individual choice, thereby eliminating any possible operation of allocation bias. Three examples may be used to illustrate the approach.

First, the effects of the Dutch famine in World War II were shown to lead to a higher frequency of central nervous system (CNS) congenital anomalies (Stein *et al.* 1975). Using this as a starting point, the same sample was used to determine whether prenatal famine increased risk for schizophrenia. It was found that it did (Susser *et al.* 1996) and a somewhat similar exposure to famine in China (St. Clair *et al.* 2005) replicated the findings. Of course, these findings do not mean that schizophrenia is ordinarily caused by prenatal famine. Rather, the key question is whether such an experience might lead to some change in the organism that could operate much more broadly. McClellan, Susser & King (2006) have suggested that *de novo* mutations induced by folate deficiency might constitute such a mediating mechanism.

Second, Costello *et al.* (2003) seized the opportunity provided by the setting up of a casino on an American Indian reservation to examine whether the relief of poverty was associated with an effect on childhood psychopathology. The experiment was possible because federal law in the United States required that a particular proportion of the profits from the casino had to be distributed to all living on the reservation without any actions by the individuals. It was possible to study change over time within individuals by virtue of the fact that the timing of the setting up of the casino came in the middle of a prospective longitudinal study undertaken by Costello and her colleagues. The results showed that the casino profits had indeed resulted in a substantial reduction in poverty and that this was followed by a reduction in the rate of some (but not all) kinds of child psychopathology.

More detailed analyses indicated that the benefits were likely to have been mediated by changes in the family.

A third example also involves the study of the removal of risk – in this case testing the hypothesis that the measles, mumps and rubella vaccine (MMR) was responsible for such a major effect in the causation of autism that it had led to a virtual epidemic associated with a markedly rising rate in the diagnosis of this condition. The experiment was possible because, at the time the rest of the world was continuing to use MMR, Japan stopped usage. The findings showed that the withdrawal of MMR had no effect on the rising trajectory in the rare of autism spectrum disorders (Honda *et al.* 2005). A similar strategy, with similar negative results, was evident in the withdrawal of thimerosal (a mercury preservative) from vaccines in Scandinavia during the early 1990s (Madsen *et al.* 2003; Atladóttir *et al.* 2007).

5.5.4 Within individual change

A further strategy that may help test a causal inference is the examination of the timing of within individual change in relation to the timing of some measured environmental exposure of a risky or protective nature (see Rutter, 2007 b). Strictly speaking, this does not constitute a true 'natural experiment', but it uses the same basic thinking and tackles the same types of question. The key methodological issue is whether the associations found reflect social selection (i.e. allocation bias) or social influence.

For example, in the field of antisocial behaviour numerous studies have shown that gang members tend to commit serious and violent offences at a high frequency. The query is whether this is because individuals with a greater antisocial liability are likely to join a gang or because there is a deviant socialisation effect of gang membership on delinquent activities.

Thornberry, Krohn, Lizotte & Chard-Wiershem (1993) used longitudinal data to make this contrast. Between group comparisons were first made between non-gang members, transient group members, and stable gang members. Then within individual over time comparisons were made to examine whether delinquent acts varied between the time before joining a gang, the time in the gang and the time after leaving the gang. For transient gang members there was no evidence of a selection effect but substantial evidence of social facilitation (i.e. crime rates were higher during the period of gang membership).

A comparable issue arises with the repeated finding that married men are less likely than unmarried men to engage in crime (see Sampson & Laub 1993). Given the fact that marriages break down, as well as being made, Sampson, Laub and Wimer (2006) tackled the problem by determining whether, over time, individuals were less likely to engage in crime during their married phase than during their earlier or later periods of not being married. Ten individual specific and ten family features were used to assess selection into marriage and these were entered into a model that also considered the variations in both crime and marriage with age (plus other time varying covariates). An inverse probability of treatment weighting (IPTW) method was used to create, in effect, a pseudo-population of weighted replicates that allowed comparisons of married and unmarried status without the need for making distributional assumptions about counterfactuals. They found an average reduction of about 35% associated with marriage – an estimate that did not vary much with whether the time interval being considered was very long (from 17 to 70 years of age) or medium (from 17 to 32 years of age).
As with other non-experimental approaches, it was not possible entirely to rule out the operation of some unconceptualised and unmeasured confounder, but it seems unlikely that any would be powerful enough to eliminate the marriage effect.

Both of these examples of within individual change concern antisocial behaviour rather

than a medical disorder but the strategy is clearly applicable to time varying exposures to risk factors that could play a part in the causal processes of disease.

5.5.5 Overview of natural experiments

No one of these designs overcomes all the problems in moving from non-experimental observation to causal inference, but, taken together, they can do much to increase the plausibility of a causal inference.

For example, the fact that several of these 'natural experiment' designs have produced evidence on environmentally mediated effects of child maltreatment on mental disorder means that some confidence can be placed on the causal inference. Conversely, the fact that all designs have failed to confirm an environmentally mediated effect of early drinking on later alcoholism means that it is unlikely that this represents causation (see Rutter, 2007 b, for details).

It is worth briefly noting what each of these various designs achieves. First the several genetically sensitive strategies do a good job in dealing with the possibility that the putative causal factor, although defined in terms of an environmental feature, actually has a causal influence by means of genetic mediation. This is a real possibility when individual variations in exposure to the causal factor involve the influence of human behaviour. This would apply, for example, to factors such as smoking, child abuse, poverty, dietary variations and family discord/conflict. Although not dealing with genetics directly, the use of an external instrumental variable such as early puberty or the use of a migration design does much the same.

Designs focusing on samples in which the total population either suffered the risk experience (as in the famine studies), or benefited from the removal of risk (as in the casino study or the stopping of use of the MMR vaccine or a thimerosal preservative) create an important experimental opportunity because they remove the element of choice, and hence eliminate selection/allocation bias (the main value of an RCT). The value of these several 'natural experiments' depends greatly on the sources of bias that need to be dealt with, and with the researchers' skill in considering possible biases and dealing with them appropriately.

The strengths and limitations of the different forms of natural experiment are discussed more fully in Rutter (2007 b). Particularly when several strategies can be used, they can help support or weaken the causal inference. Their cumulative strength lies in their power to disprove causal inferences, rather than prove them absolutely (which is not possible) – see Platt 1964, Popper 1959.

5.6 What is the place of RCTs in research into causes?

There are several reasons why RCTs constitute a strong design. They ensure that alternative causes are not confounded with the treatment conditions; they reduce the plausibility of threats to validity by distributing them randomly over conditions; they equate groups on the expected value of all measured and unmeasured variables at pretest; they allow accurate modelling of the selection process; and they allow computation of a valid estimate of error variance that is unrelated to treatment (Shadish, Cook & Campbell 2002). No other design does all of that as well. For these reasons we conclude that RCTs constitute the best way of evaluating any new treatment or any preventive intervention. It is still a concern that some people continue to be reluctant to use an RCT even when it is clearly feasible and ethical. We urge that this reluctance be overcome if there is genuine uncertainty over the efficacy of any intervention. We note, however, that there are circumstances in which it may be necessary to randomise at a group, rather than individual, level. Also, we point out that RD designs provide an alternative that has many strengths that parallel those of RCTs, even though there are more tricky assumptions (see Section 3.3).

For all their strengths, RCTs also have weaknesses with respect to the internal validity of findings (meaning the extent to which the RCT truly provides clinching evidence of causation within the sample studied – see Section 3.2). The most important of these are (Shadish, Cook and Campbell 2002; Heckman, in press; Heckman & Smith 1995):

- Failure to implement the full interventions (because of lack of adherence).
- Participants assigned to the control condition seeking treatment elsewhere.
- Treatment diffusion or dilution.
- Differential post-assignment attrition.

These problems applied to some extent in the large and rightly influential Women's Health Initiative RCT of HRT (see Section 6.3.1). Thus, 42% in the HRT group ceased to take HRT and 11% in the placebo group switched to active HRT.

The diffusion effect is probably most likely with psychological or social interventions in which there is the possibility of the comparison interventions coming to incorporate some of the elements that are supposed to be restricted to the experimental intervention. The main problem with respect to attrition is not the overall level of drop out but rather a rate of drop out or drop out mechanisms that differs between the experimental and control groups. In other words, if the level of attrition is the same in the groups being compared, but the reason for attrition is different, this may create an important bias. If that is serious, it will threaten the internal validity of the RCT.

The greater limitation of RCTs, however, concerns their *external validity*. As Cartwright (2007) has argued, RCTs provide *clinching* evidence of causation within the single study provided that the necessary assumptions are met. But if the goal is identification of a more general causal inference, there is the inevitable cost that is inherent in the assumptions – because typically they tend to be very restrictive, and the causal inference can only be strong with respect to the treatment difference that was randomised. In other words, the cost of very high internal validity may often be rather limited external validity. By contrast, non-experimental designs have the reverse set of qualities; they merely *vouch* for the causal inference but they are broad in their range of application. Hence, in many circumstances, there are many advantages to using a combination of different designs.

But how great is the generalisation concern? Egger, Davey-Smith and Sterne (2002) considered the issue in relation to the numerous RCTs undertaken to evaluate the efficacy of beta-blockers in reducing mortality after a myocardial infarction. They found risk ratios that varied across a range extending from 0.46 to 1.79. In most cases the risk ratios were in the same direction, the variation being mainly in the size of effect. The fact of the variation, however, is a reminder that, even with RCTs, no single study on its own should be regarded as definitive.

Meta-analyses may be helpful in these circumstances. They constitute a standardised means of combining data across studies in a way that weights for sample size and which tests for homogeneity of effect sizes (Shadish, Cook & Campbell 2002). Inevitably, they rely on what may be uncertain inferences regarding the comparability of studies (with respect to samples, measures, duration of follow up and interventions). It is notable that when several meta-analyses have been undertaken, they may differ in their conclusions because of differences in inclusion and exclusion criteria. Publication bias (i.e. the marked tendency for scientific journals to be much more likely to accept papers with a positive finding than those with a negative finding) is also an ever present problem.

It cannot be assumed that all variations among RCTs are due to random error, because occasionally the variation may reflect systematic variation in the circumstances associated with efficacy. Egger, Davey-Smith

& Sterne (2002) gave the example of BCG vaccination in which the risk ratios across RCTs varied from 0.20 to 1.56, but with the efficacy notably more effective in cooler climates compared with hot ones.

When the effects of treatment are likely to be modest (this is very frequently the case), large-scale RCTs are almost always required because of their strict control over both bias and random error. On the other hand, large-scale non-experimental studies (although not replacing RCTs) can contribute valuable complementary evidence about large adverse effects of treatment on infrequent outcomes when they are not likely to be associated with the indications for (or contradictions to) the treatment of interest. The qualifier is important because it is that feature that makes selection bias unlikely (Collins & MacMahon 2001; MacMahon & Collins 2001; Collins & MacMahon 2007; Vandenbrouke 2004).
So far as this report is concerned, however, the crucial restriction does not lie in any concerns over the value of RCTs but, rather, comes from the infrequency with which they are feasible in the study of possible environmental causes of disease. It is for that reason that the main focus has to be on non-experimental methods.

Nevertheless, it is important to emphasise that in many circumstances there is much to be said for using both RCTs and non-experimental methods to study the effects of interventions (which may be relevant in the identification of causes). That is because the two methods have rather different sets of strengths and limitations (Cartwright 2007; Collins & MacMahon 2007; MacMahon & Collins 2001). RCTs, unlike non-experimental methods, are able to provide clinching evidence of a true causal effect in the sample studied. In Chapter 6 we give several examples in which this clinching evidence is crucial. RCTs may be limited by a narrow focus on just one specific causal possibility but it is quite possible to compare and contrast two or more alternative causal possibilities (see Section 6.2.6 for the example of the contrast between folic acid and non-folic acid multivitamins in the prevention of neural tube defects). RCTs also tend to focus on short-term effects whereas observational non-experimental studies can assess long-term consequences both intended and unintended. Because of that, they serve a valuable role in spotting unexpected effects.

6 Examples of non-experimental research

6.1 Introduction to examples of non-experimental research

There have been many non-experimental studies designed to identify possible environmental causes of disease. Rather than seek to review these studies as a whole, we have selected three groups that differ with respect to the strength of the causal inference that is possible. First we discuss ten examples that have led to relatively strong causal inferences that appear to be justified by the evidence currently available. They vary in the strength of the causal effect and they vary in the types of evidence that point to the likely validity of the causal inference. Second, we discuss four examples of non-experimental research that have given rise to probably valid causal inferences, but for which the evidence is not quite as secure as in the first group. Third, we consider six examples in which the causal claims seem to be mistaken. With each group, we seek to identify the lessons that may be drawn.

6.2 Non-experimental research that has led to relatively strong inferences

6.2.1 Smoking and lung cancer

The effects of smoking on lung cancer (Doll & Hill 1950; Doll *et al.* 2004) constitute an obvious example of non-experimental data that has led to a causal inference that has held up. The initial finding was based on a case-control comparison in which the effect of smoking on lung cancer was very large. The plausibility of the causal inference was strengthened by the fact that it was difficult to think of an allocation bias that was plausible. Also, the large size of the effect meant that the bias would have to be huge to account for the finding (Cornfield *et al.* 1959; Doll *et al.* 2004).

Nevertheless, it took quite a long time for doubts to recede. The causal inference was much strengthened by the evidence from follow up studies that indicated the risk of lung cancer was markedly reduced if individuals gave up smoking. Although the individuals chose to give up, it is not likely that they did so for reasons that involved a lower risk of lung cancer. The biological plausibility of an environmentally mediated effect was much strengthened by animal studies that showed the carcinogenic effect of the tars involved in cigarette smoking. Further evidence has been provided more recently demonstrating that both smoking and cessation of smoking had effects on gene expression (Spira *et al.* 2004).

It should be noted that there are well demonstrated genetic influences on the liability to smoke cigarettes (see Eysenck 1980 & 1991; Fisher 1958 a & b) so genetic mediation was not impossible. Nevertheless, the overall pattern of findings indicated that it was implausible that the effects could be accounted for by genetic mediation. The genetic effects applied to the liability to smoke and not to a shared liability between smoking and lung cancer (or other adverse health outcomes). The totality of the evidence is too extensive to review here, but it is now clear that the causal inference is well justified (Office of the US Surgeon General 2004).

6.2.2 Lipids and coronary artery disease

The testing of a possible causal link between lipids, atherosclerosis (the hardening of the arteries) and coronary artery disease is unusual in that the beginnings lay in an animal study, rather than an observation in humans (Steinberg 2004; 2005 a & b; 2006 a & b). An experimental pathologist in 1913 showed that feeding rabbits very high doses of purified cholesterol in sunflower oil induced vascular lesions closely resembling human atherosclerosis both grossly and microscopically. This early research was ignored partly because rabbits are herbivores with a near zero cholesterol intake, partly because the results in rats and dogs were different,

but, probably most of all, because the findings were inconsistent with the prevailing view of atherosclerosis as a disease of ageing. During the 1940s and 1950s human experimental work provided much new information about lipoproteins in the blood and in 1956 a multi-site longitudinal study showed that lipoprotein levels strongly predicted later cardiac events.

By the late 1960s the evidence convinced many people that cholesterol levels played a causal role in atheroma. By then, studies in other animal species (including guinea pigs, goats, hens, and non-human primates) had confirmed the previously rejected rabbit findings. Despite this growing body of evidence, the lipid hypothesis had very powerful opponents who dismissed the whole idea. Proponents of the lipid hypothesis drew attention to the observational findings of familial hypercholesterolaemia (a monogenic disorder), which showed the crucial link with the gene influencing lipid levels, and demonstrated the strong association with atherosclerosis.

Ecological studies (Keys 1980) similarly showed a link between cholesterol levels and coronary deaths. The large-scale Framingham epidemiological longitudinal study (Wilson *et al.* 1980) showed that cholesterol levels are a strong risk factor for a heart attack. Dietary studies showed the benefits of reducing cholesterol intake but the benefits were small and not very convincing with respect to the causal inference. Scepticism continued!

The situation changed during the 1970s with the immense gains in the understanding of lipoproteins deriving from basic laboratory research. Brown and Goldstein went on to gain a Nobel Prize for their pioneering of this enterprise. A pathogenic model of how LDL penetrated the artery wall to give rise to the diagnostic lesion established a mechanism for the lipid hypothesis.

The US Coronary Primary Prevention Trial in the 1980s provided better evidence on the benefits of reducing levels of low density lipoprotein (LDL), a type of lipoprotein, but it was only with the development of the statins in the 1990s that a series of large-scale RCTs firmly settled the cholesterol controversy. It was not just the use of RCTs but the availability of drugs that had a very large effect on cholesterol levels that made the difference.

The lesson stemming from this nine decade story is not that RCTs provide the clinching evidence, but rather that progress came in a series of steps bringing together the crucial non-experimental epidemiological evidence, experimental animal studies, genetic evidence, clinical observations and clinical trial data. It was also critical that basic science provided key evidence on the biological mediation.

6.2.3 Perinatal studies in HIV infection

Non-experimental studies designed to address specific questions contributed to the understanding of perinatally-acquired HIV infection and have led to interventions to reduce mother to child transmission of infection from vaginal delivery and breast feeding. Much of the success of the non-experimental study approach has resulted from a well organised international collaboration devising and adopting protocols with agreed definitions and core information to address common issues across studies in different geographical situations (Dabis *et al.* 1993).

In the early 1990s, as part of the European Collaborative Study (ECS), an analysis based on 1,254 mother child pairs examined the effects of delivery on transmission risks, allowing for potential confounding factors associated with transmission (ECS 1994). The findings led to an estimate that caesarean section halved the rate of transmission. The women undergoing caesarean section differed from those with vaginal deliveries in that they were more likely to have advanced disease. This allocation bias could accordingly result in an underestimation of the protective effects of caesarean delivery. Similar results were found in other studies and

a large meta-analysis of individual data from 8,653 mother to child pairs from 15 American and European non-experimental studies reported more than a 50% reduction in mother to child transmission of HIV, independent of the use of prophylactic zidovudine. Transmission rates were 8.2% in the caesarean group and 16.7% in the group undergoing other modes of delivery. An RCT showed a similar finding (The European Mode of Delivery Collaboration 1999).

In 1985, case reports implicated breast feeding as a source of infection for infants in women infected postnatally. Infants exposed to breast feeding in these circumstances were considered at higher risk of acquiring infection due to the viraemia associated with a primary infection. There was, as yet, no evidence that women with an established infection before birth transmitted their infections to their offspring. Subsequently, collaborative studies showed an increased risk of a mother to child transmission in HIV women who breast fed their infants, and a meta-analysis confirmed the reality of the effects (Dunn *et al.* 1992). An RCT produced similar findings (Nduati *et al.* 2000).

What is distinctive about these two examples of risks to children associated with maternal HIV is that great attention was paid to methodological issues and to examining effects in different countries using comparable methods. Also, careful attention was paid to alternative explanations of findings. The iterative approach in which the non-experimental findings were put to the test in a series of studies made causal inference sufficiently plausible that policy changes followed. RCTs were confirmatory.

6.2.4 Male circumcision and HIV

During the 1980s and 1990s there were seven prospective studies and 38 cross-sectional studies examining associations between HIV and whether men were circumcised (Halperin & Bailey 1999). With very few exceptions, the risks for HIV were found to be markedly raised in uncircumcised men (with risk ratios mainly in the three to four range). The samples include two in the USA, one in India, one in Tanzania, and three in Kenya. The consistency in findings and the strength of effects pointed to the likelihood that a causal inference was justified – particularly as the inner mucosal surface of the foreskin (removed in circumcision) provided a vulnerable port of entry for HIV and other pathogens (Halperin & Bailey 1999; Moses, Bailey & Ronald 1998; Szabo & Short 2000). It had also been observed that uncircumcised men had a greater risk of ulcerative sexually transmitted diseases. An additional datum was that HIV seroprevalence tended to be substantially higher in countries with a low proportion of men circumcised. Caution was called for, however, because of the inevitable confound of variations in the men's sexual behaviour, in spite of the fact that statistical account had been taken of these in the best studies.

That the causal inference is justified has now been shown by the findings of three RCTs in three different countries, examining the effects of circumcision in early adult life on seronegative, sexually active men (Auvert *et al.* 2005; Bailey *et al.* 2007; Gray *et al.* 2007). All three showed that the relative risk was approximately halved in the circumcised group, before and after taking account of variations in sexual behaviour. With respect to the possible use of circumcision as a preventive measure against HIV in sub-Saharan Africa, it needs to be appreciated that the operations in the RCTs were performed by trained personnel using sterile equipment in appropriately equipped facilities. Surgical complications would be likely to be much greater under less satisfactory conditions (Landovitz 2007). In all parts of the world, because circumcision provides only very partial protection, there is the danger that risks would increase if the operation led the men to engage in more sexually risky behaviour. Accordingly, circumcision needs to be considered as just one element in preventive programmes, despite the fact that it demonstrably lowers the risk of HIV.

6.2.5 Blood transfusion and variant Creutzfeldt-Jacob disease (vCJD)

A further example of non-experimental research that led to a relatively strong causal inference is provided by the transmission of vCJD by blood transfusion (Llewelyn *et al.* 2004; Peden *et al.* 2004; Wroe *et al.* 2006). The number of cases identified remains tiny but the extreme rarity of the disease in the absence of transfusion risk, plus the supportive evidence from biological studies, makes the causal inference sufficiently probable to justify public health action.

6.2.6 Folic acid and neural tube defects

Low levels of folate (a B group vitamin) in early pregnancy can lead to inadequate closure of the embryonic neural tube, manifest as congenital abnormalities including anencephaly, spina bifida or encephalocele. These relatively rare malformations, collectively known as neural tube defects (NTDs), are usually serious, leading to fetal or perinatal death in some cases, or long-term disability in others.

The risk of NTDs has been reduced in many countries by the use of maternal folic acid supplementation immediately before pregnancy, and by the universal public health intervention of fortification of flour and other grain products with folic acid. The US was the first country to adopt mandatory fortification of grain products, and, since 1997, a wide range of foods has been routinely fortified. In the UK, the Food Standards Agency recommends similar fortification of most flours. The story of the discovery of the importance of folic acid in early pregnancy for the healthy development of the fetus illustrates the usefulness of non-experimental epidemiological studies, including a natural experiment, in establishing evidence to support intervention.

In the 1960s Hibbard & Smithells (1965) noted that NTDs were much more common in poorer families. This pointed to an environmental cause. Hibbard and Smithells focused on folate because in their case-control study a test that reflected folate metabolism was abnormal in pregnancies associated with congenital malformations (including NTDs). The role of social class gradient was confirmed later by Smithells and colleagues (1980) who conducted a study in which women who had a pregnancy with a NTD were given a multivitamin pill; a low recurrence rate of NTD was found.

Elwood & Nevin (1973) studied time trends and maternal characteristics in a series of 360 NTDs that arose from 41,351 births in Belfast between 1964 and 1968. Once again, there were few effects other than a marked social class gradient with NTDs being more common in poorer families.

Use of a natural experiment provided further clues. The Dutch Hunger Winter refers to the severe and prolonged famine in the Western Netherlands that occurred over the winter of 1944-45 as a result of the combined effects of a blockade by the occupying Nazi forces and a particularly severe winter causing flood, crop failure and freezing of water transport routes. The whole population was affected with little (though some) scope for personal characteristics, behaviour or choice attenuating the exposure to severe dietary restriction.

Stein and colleagues (1975) followed up the offspring exposed *in utero* to the effects of famine and maternal starvation using routine data and previously collected assessments. They created historical birth cohorts that varied by prenatal exposure to famine and the timing of that exposure during gestation; males were followed up to military induction at 18 years of age. Again, there were few health effects manifest over the first two decades of life other than a remarkable excess of congenital abnormalities of the CNS, including NTDs, stillbirths and neonatal deaths in the cohort that was exposed to the height of the famine during early gestation. Privation and timing were both important.

Subsequently, different non-experimental designs led to contrary results. Mill and colleagues (1989) undertook a large case-control study comparing the recalled use of periconception vitamin supplement use by mothers of 571 babies with an NTD, 546 with other congenital abnormalities and 573 normal deliveries. There were scarcely any effects, and no evidence of a relation between supplement use and NTDs. However, a large cohort study (Milunsky *et al.* 1989) published in the same year, involving over 23,000 births, suggested a large and specific benefit of maternal use of folic acid supplement during the first six weeks of pregnancy. No supplements, multivitamins without folate, or folate use later in pregnancy were all associated with a four-fold prevalence of NTDs in offspring compared with early folic acid supplementation.

The results of all these non-experimental studies could reflect causation (such that a lack of crucial vitamins led to NTDs) or they could reflect confounding (such that high social class women at low risk were more likely to take vitamins). Only an RCT could resolve this and in the 1980s such a trial was conducted. The MRC Vitamin Study (MRC Vitamin Study Research Group 1991) randomised 1817 women who had previously had a child with a NTD to one of five dietary conditions as they prepared for a future pregnancy: no supplementation, folate supplements, supplements without folic acid, multivitamin, and a combination of folic acid and multivitamin. From 1195 completed informative pregnancies, there were 27 with an NTD, six in the two groups with folic acid supplements, and 21 from the other two groups. Folic acid had prevented about 75% of NTDs – a result that was highly significant.

This trial, building on the foundations laid using observational designs, set the scene both for the recommendation of folic acid supplementation in high risk populations and, because of its ease and likely safety, for public health approaches and universal intervention. A recent meta-analysis confirmed the effect in subsequent RCTs (Husan & Bhutta 2007). The full biological explanation of the mechanism of the protective effect, and the interaction between genetic and environmental effects, remains obscure.

6.2.7 Fetal alcohol syndrome

Reports of possible damaging effects on the fetus stemming from mothers' heavy drinking of alcohol in early pregnancy go back very many years (see Randall 2001). However, the first postulation of a distinctive syndrome came from the observation of particular patterns of malformation in the offspring of chronic alcoholic mothers (Jones *et al.* 1973). A distinctive set of facial features was described and it was argued that these reflected damage to the developing brain brought about by the exposure of a fetus to high alcohol levels in early pregnancy. Numerous reports in humans confirmed the observation and also showed that this was accompanied by abnormalities in behavioural development (see Gray & Henderson 2006). The consistency of the observation was persuasive but the causal inference was made more likely by the link with the particular timing in early pregnancy.

What seemed to clinch the causal inference in the case of the fetal alcohol syndrome was the demonstration in studies of mice that embryos developed craniofacial malformations closely resembling those seen in the human fetal alcohol syndrome (Sulik *et al.* 1981). Subsequently, mouse studies have done much more to examine the adverse consequences resulting from fetal exposure to alcohol (Becker *et al.* 1996), and these have also shown strain differences in vulnerability to alcohol exposure and pointers to the likelihood that developmental exposure to alcohol involves reprogramming of genetic networks (Green *et al.* 2007). In short, the combination of specificity of timing and the reproducing of very comparable effects in animal models has made the causal inference indisputable.

On the other hand, many questions remain (Gray & Henderson 2006). In particular, although it is clear that alcohol exposure in early pregnancy can lead to adverse developmental effects even when the characteristic physical signs are absent, there is continuing uncertainty on how to diagnose these more subtle fetal alcohol effects. Questions remain, too, on whether the damage mainly reflects very high, but episodic, alcohol exposure (such as through binge drinking) or whether it reflects an overall 'dosage' of alcohol exposure during the key time period. Uncertainties similarly remain on the extent to which the effects extend throughout later periods of the pregnancy (albeit less clearly) and controversies remain on whether or not even low levels of alcohol exposure in early pregnancy bring about adverse effects on brain development.

One of the problems in defining the limits of the fetal alcohol syndrome in the absence of the characteristic stigmata has been the difficulty of clearly separating prenatal from postnatal effects, given that fetal alcohol exposure is most often apparent in children born to chronic alcoholic mothers – so that the postnatal environment also carries multiple risks. Studies of children exposed *in utero* to alcohol but adopted or fostered in early infancy have helped to indicate the reality of the prenatal effects (Moe 2002; Singer *et al.* 2004).

6.2.8 Rubella, thalidomide and teratogenic effects

Rubella was recognised as a distinct disease in 1881. The teratogenic effects (leading to malformations) on the fetus of maternal rubella in pregnancy were first noted only 60 years later by Gregg (1941), an Australian ophthalmologist who drew attention to the presence of cataracts, microphthalmia and a characteristic 'salt and pepper' retinopathy. A high percentage of affected infants also had cardiac anomalies and failed to thrive; in due course many were also found to have a severe perceptual deafness. Although not immediately universally accepted, other retrospective studies in a range of countries confirmed his finding (reviewed by Hanshaw *et al.* 1985).

Because the early studies had infants with congenital anomalies as their starting point, there was a risk of overestimating the rate of such anomalies. Prospective studies in the 1950s and 1960s produced lower estimates, albeit still fairly high (10%–54% - Hanshaw *et al.* 1985). These studies, however, may have underestimated the incidence of anomalies because laboratory confirmation of the diagnosis was not then possible. The situation changed in 1962 with the isolation of the rubella virus (Parkman *et al.* 1962; Weller & Neva 1962). Serological diagnostic tests thereby became possible.

During the next two years there were very extensive rubella epidemics in both Europe and North America. It was estimated that some 20,000 to 30,000 rubella damaged babies were born. It was shown that when there was virologically confirmed rubella in the first trimester of the pregnancy, the fetus was almost invariably affected and some 80% to 85% of the infants were damaged. Attenuated rubella vaccines were introduced in 1969-70 and by 2002 the majority of countries included rubella vaccination in national immunisation programmes.

The causal inference was probable on the basis of the strength of the association and the unusual nature of the sequelae. However, prospective studies were needed to confirm the strength of effect and a laboratory diagnosis provided the clinching next step leading to preventive immunisation programmes.

The association between maternal thalidomide use during the early pregnancy and phocomelia (and other limb reduction malformations) first observed by McBride (1961) and Lenz (1962) provides a similar story (Millen 1962). The strength of the effect, its very unusual and characteristic nature, and the rarity of

its occurrence in the absence of thalidomide exposure provided strong evidence of causation. In these examples, biological plausibility was clear in view of knowledge on the reality of teratogenic effects. The issue was simply whether the strength and nature of the associations in these instances were sufficient to make the causal inference. We conclude that they were.

6.2.9 Physical and sexual abuse of children

Numerous non-experimental studies have shown quite strong associations between child maltreatment – both physical and sexual – and various forms of mental disorder in childhood, adolescence and adult life.

Because the association was strong, and because it was found in all populations in which it had been studied, the possibility of this being a chance association was small. Because the maltreatment in early childhood long preceded the mental disorders with which it was associated in later life, direct reverse causation was also unlikely. On the other hand, it was certainly possible that the maltreatment had been provoked by children's difficult behaviour in early childhood. It was also possible that it reflected a shared genetic liability.

Genetically sensitive 'natural experiments' had been important in showing that environmental mediation was highly probable. Thus, as already noted, discordant twin designs showed that the mental disorders in adult life were much more likely to arise in the twin who suffered sexual abuse than the twin who did not. Moreover, the strength of association was very comparable to that found in the population as a whole (Kendler & Prescott 2006).

Multivariate twin analyses (Jaffee *et al.* 2004) were also crucial in showing a marked difference between corporal punishment (where there was no evidence of environmental mediation) and maltreatment (where there was strong evidence of environmental mediation). Biological studies have been important in showing lasting neuroendocrine effects (Gunnar & Vasquez 2006). That provides biological plausibility but, so far, it has not been shown whether the neuroendocrine effects mediate the adverse psychological outcomes. Animal studies have also shown neuroendocrine effects, although these mainly apply to the experience of neglect (which overlaps considerably with maltreatment) and chronic stress experiences.

It is noteworthy that, despite the reasonably good evidence that a causal inference is justified, the effects are diagnostically non-specific. Ordinarily, that would give rise to caution on the causal inference but, in this case, it has not really cast significant doubt on the causal inference, because the likely mediators have widespread effects on risk. It is the use of natural experiments and demonstration of neuroendocrine effects in both humans and animals that makes the causal inference likely to be justified.

6.2.10 Institutional care and disinhibited attachment disorders

For over half a century there have been observations that children in depriving institutions show maladaptive forms of behaviour. The causal inference was, however, highly questionable when the association could have been influenced either by the characteristics of the children that preceded their admission to institutions and/or biases in those children who were left in institutions as against those who either returned to their biological parents or were adopted or fostered. The causal inference has now been much more firmly supported as a result of prospective studies of children from Romanian institutions who have been adopted into well functioning adoptive families in Europe or North America.

The natural experiment in this case is provided by the fact that, in the great majority of cases, the children were admitted in the first few weeks of life (before the problems in the child could have been

observable). Moreover, few (if any) children left the institution prior to the fall of the Ceauşescu regime in 1989. Furthermore there was the opportunity to examine within individual change over time in relation to an easily timed sudden transition from an extremely depriving institutional environment to a somewhat above average adoptive family rearing environment.

The findings across studies have been highly consistent and the causal inference has appeared justified (Rutter *et al.* 2007). Moreover, the association between institutional deprivation and later psychopathology was found to apply even in those children who were not subnourished (Sonuga-Barke *et al.*, submitted). It is also relevant, however, that similar findings have been evident in comparisons within Romania between institution reared and family reared children, and that an RCT of foster care in Romania has also provided evidence for a causal effect (Nelson & Jeste, in press).

6.2.11 Lessons from case studies with relatively strong causal claims

The 'success' stories span a wide range of causal influences and an equally wide range of outcomes. However, they share several common features. First, they either concerned a very large effect (as with smoking and lung cancer) or they applied to rare and unusual outcomes with distinctive features (as with the fetal alcohol syndrome or the sequelae of profound institutional deprivation or neural tube defects or vCJD). Second, detailed careful attention was paid to alternative non-causal explanations and to how to test for their possible role. Third, all made use of multiple research designs (including 'natural experiments') with complementary strengths and limitations. Thus, the smoking research included the study of reversal effects, as did the study of institutional deprivation. Furthermore, adoption and twin designs were used to check the possibility of genetic mediation (as with abuse of children). Fourth, the causal inference was tested in multiple populations that differed in their characteristics. Fifth, animal models and human experimental studies contributed support on biological processes (as with smoking, fetal alcohol syndrome, the sequelae of institutional deprivation, folic acid and HIV). It is also the case that the apparent success stories stand out in terms of the rigour of both their measurement and their statistical analyses. In no instances, did one design provide the 'clinching' proof but, in combination, they made the causal inference a compelling probability.

6.3 Non-experimental research with probably valid causal inferences

Studies do not subdivide neatly into those in which the causal inference is certainly correct and those in which it is known to be wrong. The first group of 'success' stories have all led up to a point in which the causal inference is highly likely to be correct but, inevitably, this is a judgment at a moment in time and further research could either strengthen or weaken the inference. There needs to be a greater recognition by everyone that uncertainties are inherent in medicine – with respect to both causes and interventions (see Evans, Thornton & Chalmers 2007). The existence of uncertainty, of course, constitutes the prime justification of RCTs. With respect to causes, as we have sought to illustrate, research can do a great deal to reduce the level of uncertainty, but it would be a mistake to assume that complete certainty is usually achieved.

We now discuss four topics in which the evidence on significant associations with some disease outcomes is robust, there is in each case empirical support for the causal inference, but it is not yet quite as strong as in the first group.

We then turn, in Section 6.4, to a consideration of apparently misleading claims but with the mirror image of the caveat expressed with respect to the first group of seeming

'successes'. That is, there are strong reasons for rejecting the claim that a true cause has been identified but, once more, this is a probabilistic judgment based on the evidence available up to this point in time.

6.3.1 Hormone replacement therapy and breast and uterine cancer

The Million Women Study was set up in the late 1990s to investigate the effects of specific types of hormone replacement therapy (HRT) on incident and fatal breast cancer (Million Women Study Collaborators 2003). The sample (of over a million) was based on women aged 50 to 64 years, using the UK breast screening programme; data on HRT usage were obtained by questionnaire. Data on cancers (and other disease outcomes) were obtained from the NHS Central Registers. The follow up extended over some two and a half years for breast cancer incidence and four years for mortality.

In brief, the findings showed that the relative risk for incident invasive breast cancer was not raised in past users of HRT, even among those who had used it for more than ten years. By contrast, it was doubled in current users of an oestrogen-progestagen combination, although only slightly raised (1.3) in users of oestrogen only preparations. Moreover, among current users of combined preparations, the relative risk showed a linear dose-response relationship with duration of usage (1.45 for less than a year to 2.31 for over ten years).

Previous research showed that the risk of breast cancer varies greatly by menopausal status; in order to diminish this serious confound pre- and perimenopausal women were excluded, as were those who began using HRT before the menopause. In the population as a whole, a high body mass index is associated with an increased risk of breast cancer but the relative risk associated with current HRT usage was found to be greater in thin women. Analyses were stratified by a range of possible confounders (such as family history of breast cancer and alcohol consumption). This study was not only based on an unusually large sample, but also it used a particularly thorough approach to data analysis.

The validity of the conclusions needs to be considered in relation to three different issues. First, there is the increased risk of breast cancer in current users of combined preparations. This is consistent with the results of large-scale RCTs (Chlebowski *et al.* 2003), indicating that high quality non-experimental studies can give findings that are comparable with those of RCTs when there is not a major problem of social selection/allocation bias in relation to risk for the outcome being studied. Note that this is a key difference from the study of HRT and coronary heart disease (see Section 6.4.2).

Second, there is the lesser risk associated with oestrogen only HRT in the Million Women Study. The one RCT of oestrogen alone HRT (The Women's Health Initiative Steering Committee 2004) was partially confirmatory in that it actually showed a non-significant reduction in the risk for breast cancer. Questions, therefore, remain on this issue. Nevertheless, because of the increased risk of endometrial cancer with oestrogen only preparations, the apparently smaller increase in risk for breast cancer is of limited current public health relevance for women who still have their uterus.

Third, there is the somewhat surprising finding that there was no increase in risk for breast cancer in women who used combined preparations but who were not current users. An earlier collaborative re-analysis of 51 epidemiological studies had similarly shown no increase in risk after discontinuation for more than five years, but there was some increase in risk during the first five years (Collaborative Group on Hormonal Factors in Breast Cancer 1997).

The increased risk for breast cancer associated with oestrogen-progestagen combinations has been found in both epidemiological/longitudinal studies and RCTs, and is likely to be a valid

finding. The study is included here in the 'probable' group of non-experimental studies only because of possible uncertainties regarding the risk with oestrogen only HRT, and the somewhat greater uncertainties with respect to the lack of any increase in risk associated with past (but not current) use.

The Million Women Study also examined the associations between HRT and risk for ovarian cancer (Million Women Study Collaborators 2007). The relative risk for all current users was 1.23 but only 0.97 for past users. There was no significant difference between oestrogen only and combined preparations but, unlike with breast cancer, the risks were somewhat greater for oestrogen only. The risks did vary, however, by the type of cancer histology. The findings need to be in the 'probable' category because the mechanism remains ill understood and the lack of statistical power in RCTs for this less common cancer means that they do not contribute to the evidence base.

6.3.2 Social and economic inequality and adverse health outcomes

There is an extensive literature documenting substantial associations between income and health outcomes and between occupational status and health outcomes (Adler & Rehkopf, in press; Marmot & Wilkinson 2006). These associations apply to four somewhat different propositions. First, there is the question of whether, in a given society, poverty leads to worse health in the individuals affected (the implication being that boosting the income of the poor should improve their health). Second, there is the question of whether societies that have a more unequal distribution of income have worse health in most sectors of the population and not just in those experiencing poverty (the implication being that equalising the income distribution should improve most people's health). Third, there is the proposition that high demand combined with low control at work causes ill health in adults (the implication being that reducing demands and/or increasing control should improve health). Fourth, there is the proposition that poverty has a distal risk effect such that, although it does not lead directly to ill health in children, it makes optimal parenting more difficult and thereby affects health adversely because it has an effect on proximal mediating mechanisms.

Each of these propositions has given rise to an extensive research literature, but the distinctions among the propositions have not always been made explicit. However, most reviews of these associations – based on non-experimental data – have tended to assume that because the associations are strong, robust, and well replicated, they therefore indicate causation of whatever disease outcome is being considered. However, it is obvious that other alternatives are possible (Adler & Rehkopf, in press; Mackenbach 2002). Thus poor health may lead to a fall in income and a fall in occupational level (Cartwright 2007; Case, Lubotsky, Paxson 2002): both the income and occupational level may be genetically influenced in a way that involves a shared liability with health maintenance behaviours of one kind or another. Furthermore, the associations are known to reflect, in part, both behaviours influencing health and also access to health services.

These alternative explanations, however, apply most obviously and strongly to the associations between adult social or economic circumstances and the same person's health. It is not so obvious that this would apply if the associations are between parental income/social status and the health of the children, but the association applies strongly to children's health outcomes. Moreover, the social gradient becomes greater as the children grow older and it is not accounted for by health differences at birth (e.g. as indexed by birth weight).

Genetic mediation can be considered by determining whether the impact of family income on child health applies similarly among children who are adopted and children reared by their biological parents. The findings from

the US child supplement to the large-scale National Health Interview Survey indicated that the associations were broadly similar in these two different circumstances (Case, Lubotsky & Paxson 2002). Genetic mediation, therefore, appears unlikely (although the adoptive study findings were not presented in detail).

Natural experiments, such as the casino study already mentioned, have also shown the strong likelihood of a causal effect of income on psychopathology, albeit mediated primarily through the effects on family functioning rather than directly on child behaviour. Other natural experiments, mainly focusing on adults rather than children, have included the effects of German reunification and the effects of changes in the Earned Income Tax Credit (see Adler & Rehkopf, in press). In addition, a range of statistical approaches, such as those using instrumental variables or time series analyses, have been employed.

An RCT of moving from a high poverty to a low poverty neighbourhood similarly showed significant effects on child outcomes in the short-term, although these were not so evident in the longer term (Leventhal & Brooks-Gunn 2004; Leventhal *et al.* 2005). An earlier systematic review (Connor, Rodgers & Priest 1999) identified ten RCTs but causal inferences were difficult because health outcomes were not usually studied explicitly.

The Whitehall Longitudinal Study has shown substantial health differences between those in higher status jobs and those in lower status jobs in the civil service (Marmot 2004; Marmot & Wilkinson 2006). Because the study was of individuals all of whom were in employment, and because the population did not include those in poverty, non-causal interpretations may be less likely.

The favoured explanation focuses on the low level of control in lower status jobs leading to high physiological strain and, consistent with the hypothesis, the findings showed that the association between occupational level and health diminished when adjusted for sense of job control (Marmot *et al.* 1997). It may be concluded that there is sufficient evidence to conclude that at least part of the associations between social and economic inequalities and adverse health outcomes are likely to reflect some form of environmental causation but the causal inference is not quite as strong as often assumed, if only because the mediating mechanisms have not been tested through rigorous examination of alternative possibilities.

6.3.3 Sleeping position and Sudden Infant Death Syndrome (SIDS)

The relationship between infant sleeping position and SIDS offers another example of non-experimental research that has led to policy change in the arena of child health. From 1954 to 1988 an increasing proportion of paediatric textbooks recommended frontal sleeping for infants, possibly to avoid the risk of them choking on their own vomit (Gilbert *et al.* 2005). In 1988 an overview of largely case-control studies, instigated because of rising rates of SIDS in much of the developed world over the previous 15 years, reported that prone sleeping carried a substantially increased risk (Beal 1988).

'Back to Sleep' campaigns in several countries were followed by a fall in the rate of SIDS between 50 and 70% (Gilbert *et al.* 2005). The value of understanding the relationship between sleeping position and SIDS is illustrated by the estimated 11,000 SIDS deaths in England and Wales between 1974 and 1991 attributable to harmful health advice to sleep in the prone position (Gilbert *et al.* 2005). The case of SIDS shows that the application of the findings of non-experimental studies resulted in a demonstrable change in health policy. It is included in this section on probable causation only because of the limited body of evidence supporting the causal inference.

6.3.4 Gene-environment interactions and psychopathology

There has been awareness for a long time of the possibility of gene-environment interactions (G x E) such that there is genetic moderation of environmental risk effects (see Rutter & Silberg 2002). However, the reality of important G x E with respect to psychopathology were first clearly documented by Caspi *et al.* (2002 & 2003) with respect to childhood maltreatment and effects on both antisocial behaviour and on depression. The interaction in the first instance is provided by a variant of the *MAOA* gene and in the second instance by a variant in the serotonin transporter promoter gene. The data derived from a longitudinal study with extensive systematic data from multiple informants at multiple time periods. Stringent methodological checks showed specificity of effects and the implausibility of the interactions representing gene-gene interactions rather than gene-environment interactions. Also, the interactions applied to an environmental risk factor where environmental mediation effects have been shown in other research (Rutter, Moffit & Caspi 2006).

Two main sets of findings support the likelihood that these non-experimental data do indeed reflect a causal effect. First, although there have been a few failures to replicate, the positive replications far outweigh negative ones, and a meta-analysis of the MAOA interaction has confirmed the findings (Kim-Cohen *et al.* 2006). Second, imaging data in humans have shown, in individuals without psychopathology, that there are measurable structural and functional differences in neural functioning in the brain following exposure to fearful stimuli that vary by genetic group (Hariri *et al.* 2002; Meyer-Lindenberg *et al.* 2006). The imaging data are important in showing that there are brain effects in individuals without psychopathology – making it implausible that the genetic effects apply to mental disorders as such, and the experimental paradigm has been important in showing within-individual change.

G x E and psychopathology is nevertheless included in this middle group of examples because important questions remain (discussed by Uher & McGuffin 2007). First, there is the question of how to interpret the few non-replications among a larger number of confirmations. The problem cannot be dealt with on a football score approach so that 12 'confirmations' beats three non-replications. It cannot because a question remains on why the non-replications did not find evidence of G x E. Equally, a failure to replicate cannot possibly justify rejection of positive findings. There are numerous different reasons for non-replication to consider. These include lack of statistical power, weak measurement of the environmental cause, use of samples in which there is an unusually strong genetic or environmental component that may mask gene-environment interaction, better control of synergistic G x E, a different age/gender distribution, or a better way of dealing with scaling issues (see Eaves 2006).

Note that non-replications may sometimes increase the strength of a causal inference if the initial hypothesis specified when an effect should not be found (see Rutter 1974). Until these possibilities have been adequately studied and analysed, questions remain. Moreover, identification of G x E still leaves open the need to determine the causal mechanisms involved. Nevertheless, although important challenges remain, the validity of G x E with respect to certain mental disorders seems highly probable in view of the multiple confirmations, the confirmatory findings from animal studies and from human imaging studies.

6.3.5 Lessons from examples of probably valid causal inferences

In many respects, our four examples of probably valid causal evidence are very similar to the first group of examples in which the causal inference was stronger. Nevertheless, the differences are informative. In the case of income inequality and adverse health outcomes, the evidence on the strength

and consistency of statistical associations is extremely extensive and there can be no doubt about their presence; the uncertainties concern the details of the causal inference and the mediating mechanisms involved.

The research summarised in the reviews that we cite indicates that some of the association reflects access to services, some reverse causation (i.e. the qualities of the individuals that make it likely that they will be in low status, low pay occupations), and some of the proximal effects of lifestyle differences such as smoking, high alcohol consumption, and poor diet, to mention just three possibilities. The evidence indicates that, probably, these are not sufficient to account for the whole of the statistical association, particularly with regard to possible effects on the children. The mediation by adults' sense of lack of control in work situations has substantial support, but the mediation of effects on the children remains unclear. We conclude that it is highly probable that there are causal effects but that further research is needed to test causal inferences with respect to the different health outcomes. In addition, more evidence is needed on mediating mechanisms if there are to be sound policy change implications.

A complicating feature is that social and economic inequality is a rather broad concept so that it is not clear just which feature constitutes the causal agent. Similarly, the outcomes span a rather diverse range of disease/disorder outcomes. It is not likely that the same causal effects apply equally to all. There is enough evidence to conclude that valid causal effects are involved but uncertainties on the details mean that the validity must be viewed as only probable. The sleeping position and sudden infant death (SIDS) example is quite different in that the evidence base is more limited in terms of the range of research approaches, but the association is much more specific and there are few plausible alternative explanations. Accordingly, it is reasonable that the available evidence has led to policy change in spite of the uncertainty when considered in strictly scientific terms.

The example of HRT and breast cancer is different yet again. There is a reasonable evidence base that spans several types of research, but there are a few puzzling inconsistencies in findings and the causal effect is likely to be relatively small. Nevertheless, the causal inference is probably valid and it is sufficient for it to influence policy and practice. The lesson here is that causal inferences are always more difficult to test when effects are small; also, it takes time to build up an adequate evidence base (as the smoking and lipids stories in our first group of examples clearly illustrated).

The fourth example of G x E in relation to mental disorders provides a comparable example in that there are well replicated findings and confirmatory evidence from human and animal experimental studies. The slight uncertainty remaining concerns the unresolved questions over how to interpret the few non-replications and over the mediating biological mechanisms.

6.4 Non-experimental research with probably misleading causal claims

6.4.1 The Measles Mumps Rubella vaccine

The original claim regarding the supposed risk effects for autism from use of the MMR derived from a small study of a highly selective sample in which the causal inference lacked any kind of systematic comparison or adequate consideration of alternative explanations (Wakefield *et al.* 1998). Suggestions were made on a possible biological mediation but they lacked supporting evidence. The causal inference was based on a claimed close temporal relationship between the MMR and the onset of autism, but little attention was paid to the fact that the typical age of first manifestation (as shown by studies that preceded the use of MMR) was 18 to 24

months, which happens also to be the age when MMR is usually given. Most scientists have concluded that it was a major mistake for the journal to publish a paper with an implied causal claim that was so markedly lacking any kind of supporting evidence. However, the tentative suggestion in the published paper was soon overtaken by much stronger claims in the media by the leading scientist involved, and these went even further beyond the evidence.

It was only the public health implications of the claim (it was followed by a marked drop in the take-up of MMR and an increase in cases of measles) that led to the mass of studies undertaken to test the claim. The close temporal association between MMR and autism was quickly refuted in a paper purporting to support the claim (Spitzer *et al.* 2001). The claim switched from a focus on tight timing at an individual level to a focus on the supposed causal link over a matter of years between the use of MMR and the marked rise over time in the rate of diagnosed autism. In the event, extensive further epidemiological evidence, using a variety of quasi-experimental designs, produced consistently negative findings. Most strikingly, Japanese studies showed that the withdrawal of MMR in Japan at a time when it was in widespread use in the rest of the world was associated with a continuing rise (not fall) in the rate of autism (Honda *et al.* 2005) and no change in the rate of regressive autism – the variety supposed to be associated with MMR (Uchiyama *et al.* 2007). Even the laboratory studies have not survived the test of replication (D'Souza *et al.* 2006).

The lesson here would seem to be that improper claims based on a poor non-experimental study can be refuted by much better planned non-experimental studies that took alternative possibilities seriously. It is also relevant that most of the authors of the original paper have since published a retraction (Murch *et al.* 2004), disassociating themselves from the claims of the lead scientist. Neither that scientist, nor the media, came out of the affair with any credit. The media were particularly gullible in not appreciating the extreme weakness of the evidence. It has to be said, however, that the public's acceptance of the claim was fuelled by the entirely wrong claim by spokesmen for the Government that it was known that MMR did not cause autism. That statement was no more justified than the claim that it did. The truth, in 1998, was that there was no acceptable evidence that there was a causal effect but, equally, there were no adequate studies to test the hypothesis. The problem in the case of MMR had nothing whatsoever to do with the value of non-experimental studies; rather, it concerned a reliance on poor studies and highly biased media reporting.

Some might argue that we should not have dignified such poor research by giving it a detailed consideration. We are unapologetic, however, because it illustrates well how seriously bad research on small, highly selective samples not only can be widely accepted as having identified an environmental cause, but also can have a major public health effect (in this case, a sharp drop in the take-up of MMR vaccination).

6.4.2 Hormone replacement therapy and coronary artery disease

The most quoted example of misleading conclusions from non-experimental data concerns the effects of HRT on coronary artery disease (Grodstein *et al.* 2006; Prentice *et al.* 2005 a & 2006). Various case-control studies suggested that HRT protected against heart disease (see e.g. Grodstein *et al.* 2001; Stampfer & Colditz 1991), whereas a large-scale RCT (Rossouw, Anderson, Prentice *et al.* 2002), plus other RCTs (Hemminki & McPherson 2000), suggested the reverse.

The protective finding was biologically plausible because of the marked sex difference in rates of coronary artery disease in males and females before the menopause, and because oestrogens have beneficial effects on lipid patterns (Manson & Martin 2001). However, it was highly likely that there would be major health

related lifestyle differences between the women choosing to use HRT and those not doing so, and this major selection effect alone should have made for caution before accepting the non-experimental data as showing causation.

It is clear that the original observational claim that HRT protected against coronary heart disease was misleading (Beral *et al.* 2002; Michels & Manson 2003; Petitti & Freedman 2005) but perhaps there were aspects of the non-experimental research that should have raised more questions than they did at the time. The observational studies and the RCTs agreed well on other outcomes (including pulmonary embolism and stroke with adverse effects and hip fractures with positive effects), so the discrepancy in the case of coronary artery disease is unusual (Michels & Manson 2003). It remains possible that the effects of HRT vary according to time from menopause (Rossouw *et al.* 2007) but, as the suggestion is based on non-significant differences in special subgroups, scepticism seems warranted in view of the large risk of false positives.

There are, perhaps, three main lessons. First, when selection effects involve lifestyle differences that are likely to influence the outcomes being studied, observational studies are especially open to bias. Note, however, that it is the connections between lifestyle features associated with choice and the outcome that is crucial and it cannot be assumed that choice will always bias findings. Second, both commercial considerations and strongly held clinical views carry the danger of leading to a reluctance to accept well based, but unwelcome, findings. Third, there is not just one standard way of analysing observational study findings. Particularly when study findings are contradictory, rigorous application of appropriate novel modelling methods need to be considered. Thus, when discussing Prentice, Pettinger & Anderson's (2005 b) paper on statistical issues arising in the Women's Health Initiative, Hernán, Robins and Rodriguez (2005) suggested that it was inappropriate analysis rather than unmeasured confounding that constituted the main problem.

6.4.3 Calcium channel blockers

Calcium channel blockers provide a similar sort of lesson. The risk of myocardial infarction associated with the short acting calcium channel blocker nifedipine was first raised in non-experimental studies in 1995 (Psaty *et al.* 1995). Concern soon spread to the entire class of drugs broadly called calcium antagonists (Pahor *et al.* 2000). It took almost a decade for the question to be settled experimentally with the evidence of an RCT showed that long-acting nifedipine was safe (Psaty & Furberg 2004). The problem here with the non-experimental data was that there was the likelihood of confounding by indication. That is to say, calcium channel blockers were used to treat hypertension, which is in itself a risk factor for myocardial infarction. The avoidance of that bias is where RCTs have a major advantage.

6.4.4 Caffeine in pregnancy

Some (but far from all) non-experimental case-control studies found that pregnant women with a high caffeine intake gave birth to babies with a lower birth weight than those of women with a low caffeine intake (Martin & Bracken 1987; Vlajinac *et al.* 1997). The claimed effect was biologically plausible because caffeine readily crosses the placenta and has effects on circulating catecholamines, which have effects that could reduce fetal growth.

However, three features made the causal inference highly uncertain. First, the non-experimental findings were inconsistent; second, the supposed effects were relatively small; and third, most crucially, women with a high caffeine intake in pregnancy were known to smoke more, have a higher alcohol intake, and have attained a lower level of education. An RCT between caffeinated and decaffeinated coffee intake (Bech *et al.* 2007) found no effect of caffeine on birth weight (but a possible effect in smokers). The RCT findings apply strictly to effects in the second half of pregnancy and

allow the possibility of a slight true biological interaction effect in smokers.

6.4.5 Vitamin supplements and mortality

Oxidative stress is implicated in many diseases (Bjelakovic *et al.* 2007). Reactive oxygen molecules can damage cholesterol leading to heart disease. They might also promote carcinogenesis by inducing gene mutations or dysregulating programmed cell death (apoptosis) (Bjelakovic *et al.* 2004). The human diet is a complex mix of oxidants and antioxidants (Bjelakovic *et al.* 2004). Many observational studies have indicated that a high intake of fruit and vegetables, which are rich in antioxidants, is associated with a reduced risk of some of these diseases (Bjelakovic *et al.* 2004). However, the results of observational studies and RCTs conflict on whether antioxidant supplements, such as vitamin tablets, improve or worsen health (Lawlor *et al.* 2004).

Many observational studies seemed to indicate that antioxidant supplements reduce the risk of disease (Lawlor *et al.* 2004). In contrast, RCTs showed no effect (Lawlor *et al.* 2004). In the case of gastrointestinal cancer, a meta-analysis of RCTs initially found no evidence that antioxidant supplements are protective. Moreover, a subsequent analysis went on to demonstrate that vitamins A, E and ß-carotene may actually increase mortality (Bjelakovic *et al.* 2004 2007).

The principal cause of the confusion seems to be systematic differences between those who take vitamin supplements and those who do not (Lawlor *et al.* 2004). People who take supplements tend to be healthier, but not necessarily for that reason. Non-experimental studies may therefore be measuring the influence of other confounding factors rather than the supplements. RCTs, on the other hand, balance confounding equally between the two groups under investigation. Any differences can thus be attributed to the vitamins. People choose whether to take vitamins and similar people often make similar choices. Clearly great care should be taken when using observational studies to investigate phenomena to which people can chose whether to be exposed. The case of vitamin supplements and mortality helps demonstrate that observational studies and RCTs tend to disagree principally when there is strong evidence of allocation bias; that is, when there is error caused by systematic differences between the groups under investigation. In this case, it turns out that those who chose to take vitamins were more similar in other ways than those who chose to do otherwise. This example also illustrates that, while the protective effect of vitamin supplements is biologically plausible, it is not borne out by the research evidence.

The main conclusion, however, is that the results of non-experimental studies should be treated with extreme caution when there is inconsistency in findings across studies and when there is evidence of a strong social selection/allocation bias effect. Natural experiments could have been informative but, in the event, an RCT cast serious doubt on the causal inference with respect to vitamin supplements; the negative finding, however, might well not apply to fruit and vegetables, if the benefits derive from other constituents.

6.4.6 Early alcohol use and later alcohol abuse or dependency

There are many epidemiological studies showing that unusually early alcohol use in childhood/adolescence is quite strongly associated with an increased risk for alcohol abuse or dependency in adult life (Grant & Dawson 1997). It has been widely supposed that this represents a causal predisposing influence and that steps to discourage the drinking of alcohol until the person is older might be an effective preventive intervention in relation to a serious disease – namely, alcoholism. Considerable caution, however, is called for in making the causal inference because the same evidence shows that early alcohol use is associated with other forms of disruptive and risk taking behaviour. It has also been shown that both early alcohol use

and later alcoholism are genetically influenced. Accordingly, there is a strong possibility that the two are connected by means of a shared genetic liability rather than an environmentally mediated effect of early drinking on a later disorder of alcoholism.

As it happens, four different types of natural experiment (a multivariate twin design – Kendler & Prescott 2006; a discordant twin pair design – Prescott & Kendler 1999; Mendelian randomisation – Irons *et al.*, 2007; and the use of early puberty as an instrumental variable – Caspi & Moffitt 1991; Pulkkinen *et al.* 2006; Stattin & Magnusson 1990) have all suggested that the association probably does not reflect causation; rather, the findings point to a shared genetic liability (Rutter, 2007 b). The example well illustrates the utility of natural experiments in testing causal hypotheses – in this instance with a negative conclusion.

6.4.7 Lessons from misleading claims

By far and away the main explanation of misleading claims that have not stood up to scrutiny is that they were based on small-scale weak, pilot studies that involved inadequate controls and highly specialised samples. Often, too, they were undertaken by researchers with a very limited research track record and sometimes they represented pressure groups seeking to push a particular viewpoint (see also Evans, Thornton & Chalmers 2007). It is notable that it has been estimated that about half of all presentations at conferences never appear in peer-reviewed scientific journals (Scherer & Langenberg 2007). Peer-reviews are by no means perfect as a means of establishing quality but, of the methods available, they constitute the most satisfactory first sieve. In that connection, we express concern that there is a trend in electronic open access journals to publish prior to peer-review. In the longer term, the stronger test of quality is replication by independent research groups – preferably using improved methods of measurement and analysis and using additional steps to rule out (or rule in) the likely operations of the various forms of bias. The message needs to be that everyone should be extremely cautious before accepting that findings are conclusive if preliminary studies have not gone through rigorous peer-review prior to publication. Conference proceedings and non-refereed chapters in books should be treated similarly with circumspection. The level of confidence in the causal inference should increase when several high-quality replications have been undertaken, but the confidence in the causal inference becomes reasonably strong when similar findings derive from several divergent research designs and when specific steps have been taken to deal with the forms of bias that could affect the causal claim.

In this Chapter we have noted six examples of causal claims in which the weight of evidence clearly indicates that the claims are highly likely to be wrong. They have in common a degree of supposed biological plausibility (albeit very weak in the case of MMR). However, the problem in all six instances was the strong expectation of bias stemming from selection or indication, plus in the case of MMR the involvement of litigation. There should be especial scepticism over causal claims that stem from studies having these problems.

7 Identification of causes and implications for policy and practice

7.1 How and when to act on identification of causes of disease

As noted in Chapters 1 and 2, the prime reason for seeking to identify the causes of disease is the expectation that the information should guide the design and implementation of preventive or therapeutic interventions. It is the potential for causal evidence to provide guidance on either policy change at a public health level, or alterations in professional practice in the case of individual patients that justifies the endeavour – whether this is basic science, or clinical experiments, or RCTs, or non-experimental studies. In order to decide when and how to act on causal evidence, it is necessary first to have a quantified measure of the degree of risk of strength of the causal effect. In addition, however, it is highly desirable to have an understanding of how the causal effect is mediated. We consider these two issues first.

7.2 Quantifying risk

In any study seeking to identify causes, a key issue concerns the size of effects. On the face of it, that would seem to be a straightforward matter that should be open to an easily understandable number that is not open to misinterpretation. Thus, if, say, an identified cause trebles a person's risk of some serious disease, that would appear a huge increase in risk that must have enormous public health consequences. But it is not quite as simple as it seems.

Let us take Down syndrome as a much studied condition on which good epidemiological data are available. High maternal age is a well demonstrated feature that greatly increases a woman's chance of having a baby with Down syndrome. For women over the age of 40 years the likelihood is some 16 times higher than that for women aged 20 to 25 years – i.e. there is a major increase in relative risk (Rutter 2006; Tolmie 2002). But the same data show that the absolute risk is very low. The proportion of babies with Down syndrome born to women over 40 years of age is a mere one percent. In other words, the chances of an older woman having a normal baby far outweigh the chance of a having a baby with Down syndrome – by a factor of 99 to 1. That may be reassuring to the individual woman but the huge increase in relative risk would seem to imply a large effect at a population level. But, even that does not follow. The majority of babies with Down syndrome are born to young mothers! That comes about because far more babies are born to young mothers than to older mothers. Accordingly, the population attributable risk (i.e. the absolute increase in risk due to the causal factor) is very low; that is because it is hugely influenced by the population frequency of the causal factor.

Sticking with the Down syndrome example, the attributable risk issue is shown even more strikingly by the effect of Down syndrome on IQ level. In the population as a whole the correlation between Down syndrome and IQ is exceedingly low – about 0.076 (Broman, Nichols & Kennedy 1975). One might think that this is such a weak effect that it is not worth bothering with. However, the average IQ of a person with Down syndrome is some 60 points below the general population mean – a massive causal effect at the individual level. The very low population attributable risk is simply a consequence of the relative rarity of Down syndrome in the general population.

All of this is well understood by epidemiologists, but the reporting of level of risk or size of effects by researchers all too frequently blurs these crucial distinctions. Moreover, misrepresentations of the strength of effects is even more common in reports in the media. For example, although the proportion of the population to which the risk applies may be clearly stated in the text, it is much more likely to be misleadingly presented in the headlines.

A further additional point, to return to an earlier discussion of the meaning of a cause, is that some members of the public are inclined to dismiss causal claims because they know people who did not have the predicted outcome – the aunt who was a heavy smoker all her life but yet lived to 96 years – or the heavy cannabis user who nevertheless did not develop schizophrenia. The presentation of causal effects in both the media and scientific papers needs to emphasise both the probabilistic (and non-deterministic) nature of the causal effect and also the importance of individual differences in response.

In some cases, we have a limited understanding of the causes of that variation in response (as with some instances of G x E) but in many cases they have yet to be identified. What we do know for sure is that heterogeneity in response is usual. Paling (2003) provided a helpful set of suggestions on how to help patients understand risks and we suggest that these are equally applicable to communications by researchers and by journalists. Gigerenzer (2003) specifically recommended that greatest clarity is achieved by using natural frequencies (i.e. simple counts) rather than probabilities to describe levels of risk.

7.3 Mediation of causal effects

From the perspective of public health policy or clinical practice, it is clearly highly desirable to determine the mechanisms that mediate causal effects. Without that knowledge, there is the considerable danger of focusing preventive or intervention strategies on the wrong facet of the causal factor. Strictly speaking, of course, it is possible to make a strong inference on causation without knowing just how the causal effect operates. On the other hand, the causal inference is much strengthened if its mechanism of action can be demonstrated. Equally, as we note, there needs to be substantial scepticism if there is no known mechanism by which the cause might operate; or if the hypothesised mechanism is inconsistent with existing knowledge on disease processes.

Accordingly, we argue that, if the non-experimental evidence (including the testing and rejecting of competing non-causal alternative hypotheses) is sufficiently strong, the causal inference should not be held back simply because the mediating mechanism remains uncertain. However, we also argue that research to determine the mediating mechanism should be viewed as part of the same endeavour to identify causal mechanisms, and not just some optional, secondary, later stage enterprise.

Thus, the examples of smoking and lung cancer, and the fetal alcohol syndrome, had the causal inference greatly strengthened respectively by the evidence on carcinogenic effects and teratogenic effects – in both cases through animal models. Similarly, the inference on the causal impact of institutional deprivation was bolstered by the evidence on both neuroendocrine and neural structure effects. This example well illustrates, however, the difference between demonstration of a possible relevant biological mediating mechanism and the demonstration that that mechanism does actually mediate the psychopathological sequelae (such evidence is still lacking).

7.4 Decision making on research evidence

It would be quite wrong to suppose that research evidence, in itself, can be sufficient to determine policy. It cannot. Smoking can be taken as an example to illustrate some of the key issues. To begin with it is necessary to consider the robustness of the evidence. As early as 1962, Government reports in both the UK (Royal College of Physicians 1962) and the USA (Office of the US Surgeon General 1964) pointed to the likely serious health hazards of smoking tobacco.

At that time, however, the evidence, although persuasive, was far from compelling. Possibly appropriately, therefore, it led to the provision of warnings, but nothing stronger. The situation today is quite different (Office of the US Surgeon General 2004); probabilities have become virtual certainties. That certainly applies to direct individual exposure but also (albeit to a lesser extent) it applies to involuntary passive exposure (Office of the US Surgeon General 2006). But at what point in the causal chain should the preventive action be directed?

We know that commercial advertising fosters smoking and that smoking is influenced by both costs and legislation (Chapman 1996). Accordingly, in many countries, the main interventions have focused on the commercial origins of smoking rather than individual smoking habits (although these, too, are being targeted).

Policy, however, needs to pay attention to both costs and benefits, and especially to the risks of doing something versus the risks of doing nothing. Other examples may be used to illustrate these points. It is known that lead in high dosage is a serious neurotoxin. Extensive research showed the likelihood (but not certainty) that there were lesser effects from lower level exposure and, moreover, that there was no clear cut threshold below which lead exposure was safe (Rutter & Russell Jones 1983; Schwartz 1994; Wigle & Lanphear 2005). Because there were no known benefits from the ingestion of lead, there came to be a general acceptance that it was desirable to eliminate lead based paints and to cease adding lead to petrol (gasoline) – see Rutter 1983.

The more recent example of the mercury based preservative 'thimerosal' that used to be used in vaccines, provides a similar story. As with lead, there was no doubt that high dosage of mercury led to serious neurotoxic effects and there were no known benefits. The evidence that thimerosal led to substantial and significant damage to health was decidedly weak (see Rutter 2005) but, given that it was not necessary and caused no known benefits, most governments have decided to replace thimerosal with other preservatives. In both instances, the risks of doing nothing seemed greatly to outweigh the risks of doing something, even though the scientific evidence was not strong.

Alcohol presents another parallel, but with the difference that moderate usage may have some benefits and a large majority of the population wish to be able to drink alcohol containing beverages (Academy of Medical Sciences 2004). Accordingly, despite the health risks of high alcohol consumption, it would not be sensible to attempt to prohibit it completely (the American prohibition experiment makes that clear).

On the other hand, there are at least two circumstances in which research shows clear-cut risks and no benefits – namely, driving motor vehicles when intoxicated (Academy of Medical Sciences 2004) and drinking heavily in the early months of pregnancy (Gray & Henderson 2006). Accordingly, in both cases, governmental action has been taken to deal with these specific risks. But, with respect to drinking in pregnancy, research is inconclusive on whether low levels of alcohol consumption, especially later in pregnancy, cause significant risks. At the time of writing this report, the Government has decided to issue warnings on the dangers of drinking any alcohol whilst pregnant (or intending to become pregnant). They have been explicit that this new advice is not based on new research, but rather is based on a judgement that, given that we do not know that it is ever safe and given the indications that many people are poor at judging how much they drink, this is the safest advice to give. It remains to be seen whether that was an appropriate judgment.

Folates and fluoridation provide a slightly different example. There is strong scientific evidence that folates in pregnancy protect against neural tube defects and that fluoridation protects against tooth decay.

On the other hand, there may be some minor risks of slight health disadvantages from fluoridation in a few vulnerable individuals. Should these unusual (and uncertain) individual risks prevent the use of a prophylactic that would protect millions? Exactly the same argument applies to vaccination.

We give the examples, not to urge any particular action, but rather to emphasise that value judgements, as well as scientific evidence have to be involved in policy decisions. Engagement with the public will also be desirable in circumstances in which there are ethical concerns or major uncertainties on risks and benefits. We agree with the 2006 Science and Technology Committee report that there is a greater need for openness and transparency in how the Government deals with these dilemmas. It needs to know how to ask the right questions of scientists and how to judge the quality of scientific evidence, but also it needs to be more forthrightly honest in explaining how it uses scientific evidence and why and how considered value judgements need to influence policy decisions.

Two further points need to be made. First, it should not be assumed that public opinion is fixed and impervious to change. The changed attitudes to smoking and increased acceptance of banning smoking from the workplace constitute compelling examples of changes in attitude that have made possible legislation that would have been bitterly opposed in the past. Second, even when there is very strong evidence that removing a cause of disease is desirable, it is not always so straightforward to predict the consequences of policy actions. The lesson is that any substantial policy change ought to be subjected to rigorous evaluation. That happens all too rarely.

7.5 When should identification of causes of disease lead to policy action?

In summary, six main points need to be made with respect to decision making on actions (either with respect to individual patient care or public health policies) on the basis of evidence that purports to identify a modifiable element in the causal process leading to disease. First, it is very rare for a single breakthrough study to make any causal inference certain. Rather, as both the smoking and cholesterol stories indicate well, the evidence coming from a range of different sources and using varied research designs gradually over time builds an increasingly strong causal case. Non-experimental studies have played a key role in the causal arguments but the causal inference has usually required additional input from natural experiments providing extra research leverage, from animal models, from biological understanding stemming from basic science, and (in the few circumstances in which they are possible) from RCTs. Inevitably, it is a matter of judgment to decide when the evidence is sufficient to act.

Second, both an extended follow up and different forms of statistical analysis may call for a change in the interpretation of the evidence. An example is provided by an RCT to test the effects of a welfare-to-work programme in the USA (Hotz, Imbens & Klerman 2006). The background concept was that reliance on welfare had negative effects on families and children. The RCT compared a designed welfare-to-work scheme in six Californian counties with a control group to whom such services were denied (but who could seek alternative services in the community). The findings showed a significant and sizeable effect of the programme on employment and income over a three year period in one of the counties (Riverside), less, but still significant, effects in one other, but no significant benefits in either of the other two considered in detail. The findings were interpreted as demonstrating that a labour force attachment (LFA) approach (i.e. a focus on getting everyone

into a job even if it was low paid) was superior to a human capital development (HCD) approach that provided education and vocational training in order to improve the job related skills of welfare recipients. The inference was drawn because Riverside emphasised LFA. It became a model for welfare-to-work programmes across the USA.

The inference was a shaky one at best because of major differences among the six counties in both employment opportunities and family characteristics. Also, counties were able to choose the degree to which they adopted LFA or HCD approaches. The randomisation was between any of the welfare-to-work approaches and no intervention, and not between LFA and HCD. As it turned out, a longer (nine year) follow up showed a loss of effect of the programme in Riverside, whereas the gains over time were seen in the other programmes. A more detailed regression adjustment non-experimental form of analysis (that had to rely on various assumptions), when combined with the RCT, showed significant heterogeneity in effects across counties. Putting all the findings together, the conclusion was that whereas LFA was more effective than HCD in the short term, HCD was more effective in the longer term. Also, however, the effects varied by social context.

We present the study in some detail, not because we wish to draw conclusions about the merits and demerits of LFA as compared with HCD, but, rather, to emphasise the risks in overlooking the initial design features (i.e. the RCT did not compare the two approaches), to note social contextual effects, to indicate that long-term and short-term effects may differ, and to show the possible utility of combining experimental and non-experimental methods.

The third point is quite different; namely that the type of policy change will often need to be determined by whether or not risk effects extend beyond the individual. Smoking provides the example here. The evidence that smoking causes major health risks for the individual who smokes has been persuasive for several decades, and the degree of certainty has been sufficient for governments to take steps to persuade individuals not to smoke. Nevertheless, although the risk effect was very strong, it would not necessarily justify banning smoking from public places.

The situation changed radically, however, with the growing evidence of the health damaging effects of passive smoking (Office of the US Surgeon General 2006). Before policy change could be justified, however, it was necessary to determine the degree of risk and to check whether the causal inference was justified. Numerous studies have shown carcinogens in second hand smoke – thereby showing that the agents involved in carcinogenesis were present. Moreover, it was found that exposure of non-smokers to second hand smoke caused a significant increase in urinary levels of the metabolites of such carcinogens – showing that these got into the body. A limited number of animal studies went on to demonstrate actual increases in tumour formation (although the findings are less clear cut than with active exposure). Finally, human studies showed a 20% to 30% increase in the risk of lung cancer from second hand smoke exposure stemming from living with a smoker.

Obviously, this is far less than the risk associated with active smoking but it is sufficient to have public health relevance. Similar findings applied to other health outcomes such as coronary artery disease or respiratory disease. It was this body of different types of evidence all pointing to the same conclusion that justified the bans of smoking in enclosed public places introduced in many countries.

Fourth, attention will always need to be paid to the nature and severity of the risks associated. In that connection, two issues need to be considered: the strength of the causal inference and the costs of acting versus not acting on the evidence. Both involve the need to take a decision in the context of uncertainty (see paragraph 17 of Appendix I).

Bayesian approaches seek to quantify the decision making on the basis of prior probabilities and expectable effects. A key feature of such approaches is that they aim to be able to take account of each new set of evidence as it becomes available. That is, it accepts that all decision making is provisional (because new evidence may change things), that it is crucial to determine over time whether the growing body of findings is strengthening or weakening the causal inference, and that it is helpful to quantify (so far as possible on the basis of reasonable assumptions) the relative costs and benefits of inaction versus different forms of intervention or policy change.

The risks of both action and no action may sometimes be substantial, although unequal. Thus, with respect to male circumcision and HIV, the needed intervention is a surgical procedure that causes some risk and the non-action concerns a very serious adverse health outcome. Many people would consider that the balance points to action.

Sometimes the risks of not acting far outweigh the risk of action – using statins to prevent coronary artery disease would be an example of this kind. In other cases, the balance may point in the opposite direction. Thus, the use of oestrogen only HRT seems safer than combined preparations with respect to the risk for breast cancer and coronary artery disease, but equally effective in reducing the risk of hip fractures (Women's Health Initiative Steering Committee 2004). On the other hand, the RCT indicated an increased risk of stroke, and earlier evidence had shown a substantially increased risk of endometrial uterine cancer. The risks of action seem too high in the case of women who have not had a hysterectomy and the stroke concern calls for caution even in those without their uterus. As always, a careful clinical decision analysis is needed (McPherson 2004; Minelli *et al.* 2004).

Fifth, taking similar considerations even further, there will be circumstances in which, even when there may be uncertainties over the health risks, the hazard is one that needs attention in its own right. Thus, this obviously applies in the case of child abuse and many would consider that it does too in the case of poverty and marked income inequalities.

Sixth, policy or practice interventions on the basis of evidence on identification of a causal element needs to lead on to RCTs to assess the efficacy of the intervention. The need arises from two separate, but linked, considerations. The intervention should be logically connected to the evidence on cause, but it does not follow inevitably that the intervention will achieve its objective. The campaigns to deter people from smoking and to limit alcohol consumption illustrate the point. The related concern is that the intervention should be planned in a way that will be informative on the postulated mediating mechanism. Thus, it is necessary to ask, not only whether statins are effective in lowering cholesterol levels, but also whether the health benefits are mediated only through that mechanism.

7.6 Governmental attitudes to research

As we noted in our discussion of the identification of causes, the prime incentive for research to identify causes is provided by the potential for prevention or interventions. Accordingly, we need to pay attention to how research should influence policy. We note with great concern the House of Commons Science and Technology Committee (2006) conclusion that there is both a perception that there has been a decline in scientific expertise in the civil service and an accompanying perception that scientific skills may constitute a hindrance in career progression. We lack evidence on the extent to which this perception reflects a reality but, if it does, it is a very serious concern. We note their view that the scientific advisory system in the US has some important advantages and our impression is the same.

We strongly support the Select Committee's recommendation that all senior government officials and all policymakers should have a basic understanding of scientific methods and of the importance of peer-review.

In that connection we deplore the example of the 'Sure Start' initiative in which the Government not only ruled out randomised control designs to evaluate efficacy but also insisted on a recruitment strategy for 'Sure Start' areas that severely jeopardised the possibility of a non-experimental design to test efficacy (see Rutter 2006 & 2007 c).
It is difficult to avoid the conclusion that the Government favours everyone else adopting evidence-based practice so long as they do not have to do so. The approach has been described as policy-based evidence, rather than evidence-based policy (Brown 2001). The Select Committee also stressed that research used to inform practice must be truly independent. They added, and we agree, that the current mechanisms for commissioning research are not optimal in delivering that objective (see paragraph 40). Furthermore, they argued that the Government needs to put in place incentives to encourage Departments to take a longer-term view of the utility of research in developing future policy, and not just in supporting present policy (see paragraph 44). Again, we agree.

8 Communicating the findings from causal research

In recent years, there have been several authoritative reports presenting advice and recommendations on the communication of research findings (Des Jarlais et al 2004; Royal Society 2006; Royal Institution of Great Britain, SIRC and Royal Society 2001; Social Issues Research Centre 2006), as well as similarly authoritative reports on the desirable connections between science and public policy (House of Commons Science and Technology Committee 2006). All were agreed that the prime responsibility lies with the researcher to communicate accurately, clearly and fairly what the study set out to do, how it sought to accomplish its aims and how secure were the findings, as well as the confidence that can be placed on causal conclusions, and the generalisability of the conclusions to the population at large. In relation to all these issues, the researcher should be expected to write in ways that are readily understandable to non-experts. Whereas that may not be possible within highly technical scientific journals, it is crucial that the communication should be readily understandable in writing or speaking to policymakers or practitioners, in all press releases, and in all interviews with the media.

Although many of the recommendations apply to any kind of science, in studies involving populations it is essential to be explicit on:

- The target sample or population and the criteria for recruitment.
- The sampling and the recruitment setting.
- Whether the sample was large enough for the intended purpose.
- How the postulated causal influence was conceptualised and measured.
- How the severity and duration of the causal influence was measured.
- What steps were taken to determine whether the findings applied only to certain subgroups.
- What theory or body of evidence led to the focus on that supposed causal factor.
- What steps were taken to rule out competing alternative non-causal (or different causal) explanations.
- How non-participation in the study or attrition from the sample might have biased the results and how the possibility of bias was dealt with.
- What statistical methods were used and what assumptions they relied on.
- Whether the participants were volunteers or selected at random.

The style and balance of communication is equally important. Thus, are the findings preliminary and as yet inconclusive? Have the findings been replicated in other samples? How do the findings differ from earlier research? Are there reasons why more weight should be placed on the new research? If so, what are those reasons and how solidly based are they? There is often considerable pressure from funders, employers and advocacy groups to make unduly strong claims. Whilst reasonably emphasising the importance of the science, great care should be taken to avoid exaggerating the findings. In particular, care should be taken to avoid describing the results as a 'breakthrough'. True breakthroughs in science are rare.

All of these concerns apply to the range of non-experimental studies that constitute the focus of this report. However, some have particular importance. We highlight several key issues, expressing them in the form of questions. Has the possibility of bias in sampling been examined in detail? Most crucially, is it possible that individuals who either chose a particular experience or had the opportunity to have the experience differ systematically from those not having the experience (i.e. selection or allocation bias)? Might such differences influence the disease outcome being investigated? If participation in the study was influenced by some professional decision on need (i.e. intention bias), could this influence the assessment of the hypothesised causal

effect? Might the experience of the influence being investigated increase or decrease the likelihood of the disease outcome being noticed and recorded (ascertainment bias)? These three sources of bias constitute the major sources of misleading findings from non-experimental research and the importance of measuring them in a detailed discriminatory fashion, and taking their effects into account, cannot be overestimated. This is because of the ever present hazards in drawing causal inferences from observational associations in non-experimental studies.

When communicating research about the causes of disease it is necessary to ask what were the most important non-causal alternatives that needed to be taken into account – a shared genetic liability, reverse causation, or some alternative (but associated) causal influence? Were appropriate 'natural experiments' or 'quasi-experiments' (together with appropriate analytic strategies) used to obtain greater research leverage on the causal inference? If they were, what did they show? If they were not used, what limitations should that place on the causal inference? Are there animal models or human experiments that could provide additional testing of causation? If undertaken, what did they show? If not used, will they constitute a further step in the research endeavour?

In considering the risk or protective effects found in the research, the researcher should explain carefully, using readily understood real life examples, what the implications are for the general public. If, for example, the adverse causal factor (say a medication) is associated with a doubling in risk, does this mean an increase from, say, 1 in 10 getting the disease to 2 in 10 doing so, or does it mean an increase from 1 in 10,000 to 2 in 10,000? Does it apply to any use of the medication or only to unusually regular high dosage? Does it apply to everyone equally or does it apply mainly to those at an unusually high risk for other reasons? Where possible use simple counts to describe risk, rather than probabilities.

On the whole, researchers have become rather better at explaining these implications but, in their desire to have eye catching headlines, some media are inclined to gloss over the crucial caveats and qualifications. In that connection, the general practice of having headlines written by someone other than the science reporter presents a particular problem and all media need to build in appropriate checks to avoid the serious problem of misleading headlines.

During the last few years, regulations with respect to declarations of interest (sometimes expressed as conflicts of interest) have been substantially tightened up and that is greatly to be welcomed. Nevertheless, it is important to appreciate that almost all researchers have a relevant interest. This may be in support of their employing institution, or their source of funds, or because they could make a profit out of the commercial use of some product deriving out of the research findings. The key need is transparency and honesty in the reporting of such interests. It is up to readers to decide for themselves whether or not they consider the conflict of interest to be sufficiently serious to cast doubt on the claims, but the total avoidance of relevant interests is neither practical nor desirable.

The danger and damage comes from three main sources. First, there is concealment of interests. The only safe policy is to assume that if a researcher has chosen deliberately to hide a relevant interest that is because it did create bias and s/he wishes to prevent people knowing that. Second, there is the serious bias created by the sponsors of research censoring or distorting findings or suppressing their publication. Sometimes, too, the researchers who are named as authors have been prevented from having full access to the data and for the statistical analyses. 'Ghost written' articles are still occurring and it is essential that they be outlawed. The scandals associated with such practices by tobacco companies (Collin, Lee & Gilmore 2003; Glantz *et al.* 1995; Hilts

1996; Ong & Glantz 2000) have become well known, and similar practices by some (but not most) drug companies have also received publicity. We need to appreciate, however that government departments sometimes do the same. We support the House of Commons Science and Technology Committee (2006) in their direct recognition that research must be independent from outside influence and that government must publish the evidence, analysis, and relevant papers it receives. Openness and transparency must be preserved.

As is discussed, there is a third source of bias, albeit of a different kind, that also needs to be highlighted. That is the non-publication of negative findings (Scargle 2000). Research funders are reluctant to fund replications and journals are reluctant to publish them, especially if they are negative. Similarly, employing institutions give little credit to replications. The net effect is a real bias in favour of positive findings. Of course, some negative findings never get published because they are of poor quality but some very poor positive findings do get published because they are newsworthy. Fenfluramine as a treatment for autism and MMR as a cause of autism are well known examples – if only because they were both published by high quality, high impact scientific journals. There can be no mechanical rule to prevent this problem but we recommend that the bar for acceptance of papers for publication should be higher if the biological basis is weak, if the research design is shaky, and if the public health risks, should the claim be proved false, would be serious and widespread.

Both editors of medical journals and peer-reviewers of scientific papers have dual, and occasionally conflicting, responsibilities: to ensure preservation and strengthening of the research literature, and to promote the publication and dissemination of new ideas and discoveries that may challenge established belief. A key task of any editor and peer-reviewer is to manage risk; that is, to balance newness and scientific validity. There is a strong case for journals to be more vigilant when considering for publication potentially controversial findings – especially if they might influence clinical practice or change health behaviours. A panel convened by *Science* after the recent South Korean cloning fraud concluded that research papers should be risk assessed by editors to identify particular features of a piece of research that means that it should be subject to more intensive peer-review. In extreme cases, there may be a need to make primary data available for verification of conclusions.

A high risk paper is one in which a result is likely to lead to high scientific, public health or policy controversy, and where a result gives rise to sharp disagreement among reviewers. Similarly, editors need to consider carefully when such high risk papers should be 'fast tracked' and when they should be accompanied by commentaries or critiques that highlight key issues. Sponsors, too, have a responsibility on how they conduct debates about risk, just as regulators have to decide when to take action. Inevitably, there will often be occasions when both policymakers and regulators have to act quickly despite great scientific uncertainty.

We also consider that it is desirable that editors should strengthen the collective responsibility of co-authors who should take a shared ownership of the totality of any piece of research and of the messages that are imparted in the papers. It is suggested that journals should consider sending reviews to all co-authors and to require co-authors to sign off the final accepted manuscript. At the time of publication, the editor should consider running an editorial to place the findings in context with the totality of the available evidence. Any press materials issued by the journal or the host institution should, similarly, seek balance, point to important limitations in the work, and note the issues with respect to public health implications.

With respect to non-experimental research, editors should support the creation of guidelines for such research and, when such guidelines are available and approved by reputable scientific authorities, should apply those guidelines to the papers they publish.

9 Conclusions

In concluding, we consider the six questions set out in Chapter 1 of the report:

9.1 When are causal inferences from non-experimental studies justifiable?

Although there is much to be said on the superiority of experimental methods for testing causal hypotheses, it is evident that they are neither practicable nor ethical in the case of many hypothesised causal influences on human disease. Thus, toxins, pesticides and child abuse cannot be given to volunteers in order to test whether or not they cause disease. It would be a counsel of despair to argue that, on the grounds that they are not open to experiment, such possible disease causing agents cannot be investigated and, therefore, cannot form the basis of preventive or therapeutic interventions. We noted the many important successes of non-experimental research in identifying major causes of serious disease, with the findings being taken forward in the form of crucial changes in policy and practices.

In short, we come to the firm conclusion that, if properly conducted, non-experimental studies can and do provide the basis of reasonably secure causal inferences. It is not that they ever provide 'clinching' positive proof of causation. Rather it is, in the tradition of Platt (1964) and Popper (1959) that they provide the opportunity for **disproof** of causal hypotheses. It is only when such hypotheses have survived the severe tests of possible competing non-causal alternative explanations that the causal hypothesis becomes credible. Such tests include variations in study design, contexts, sensitivity measures, reversal designs examining the effects of removal of the postulated cause, and a range of 'natural experiments' that serve to pull apart variables that ordinarily go together. In summary, causal inferences from non-experimental studies are justifiable when possible non-causal explanations have been explored thoroughly and when there is no reasonable alternative explanation for the results.

Nevertheless, it is important to recognise that not all hypotheses are created equal (Susser *et al.* 2006). The more radical the hypothesis (in the sense that it is inconsistent with generally accepted theory and previous research findings), the higher the empirical bar should be set before an association is accepted as causal. A somewhat related issue is that, so far as policy decisions are concerned, subjective judgements cannot be avoided. Such judgements will depend on what is at stake if we are wrong in either direction. There are no unambiguous rules or criteria that enable that problem to be circumvented or settled by box ticking practices, but sound scientific methods are more likely to get us to the truth more often and more quickly.

In returning to the challenge with which we started, we note that much the most important reason why causal claims quickly get refuted or overtaken is that reports derive from preliminary communications of small, often poorly conducted, studies presented at meetings of one kind or another before the research has been subject to peer-review and published in a reputable scientific journal. About half of such presentations never see the light of day in the sense of published reports that are open to critical scrutiny. The old computing adage of 'GIGO' (garbage in, garbage out) applies. We hope that journalists and their editors will exercise self-discipline in not giving these premature claims media coverage. Even with good studies, however, the key guidance in science is to believe nothing until the uncertainty is reduced through confirmatory findings from further investigations using different samples and different measures of the same basic constructs.

As ever, the golden rule is that, even with the best conducted studies, replication is essential. That applies to RCTs as well as to non-experimental studies and it constitutes the reason why the UK National Institute for Clinical Excellence (NICE) places greatest reliance on either systematic reviews of **multiple** RCTs, or meta-analyses.

Clinicians, scientists and journalists need to exercise great caution before drawing firm conclusions on the findings of just one piece of research. The great majority of false claims are not a consequence of the weakness of any one research strategy but rather reflect the vagaries created by chance on a variety of biasing features. Pooling data across samples, or meta-analyses, constitute useful means of determining consistency across studies. Note, however, that meta-analyses of RCTs are essentially treating them as observational studies (because there is not randomisation across studies). Prospective planned meta-analyses can be very informative, nevertheless, in narrowing the confidence interval. Note, too, that meta-analyses vary greatly in quality, if only because studies use different approaches and different measures. Combining data at an individual level will always be preferable to combination at a study level.

Nevertheless, there are reasons why valid findings may not always be replicated. Thus, causal effects may be dependent on context – either biological or social. As we have noted, susceptibility genes may not have their main effects on some disease/disorder outcome, but rather they may moderate environmental risk mediation (or, expressing the same point a different way, environmental effects may moderate genetic influences). Also, the effects of one genetic variant may be influenced by the presence/absence of other genes, and some environmental effects may vary by prior sensitising or steeling experiences. Although possible contextual effects are important, they need to be tested for systematically, and not just assumed.

9.2 Can non-experimental studies give rise to a causal inference?

Let us now focus on the basic question of whether non-experimental studies can ever give rise to a valid causal inference. We conclude that they can. It is not that they can **prove** causation, but rather that they can build up an ever increasingly strong case that a causal effect is highly likely, if they are supported by multiple varied research strategies that differ in their patterns of strengths and limitations, **and** if they survive multiple attempts at disproof. When these criteria are met in strong degree, non-experimental findings may give rise to a sufficiently robust causal inference to warrant policy or practice actions.

Our review of relative 'success' stories from non-experimental studies led us to identify five key characteristics. First, they have shown a large risk effect. We accept that many true risks have only a small effect but they are much more difficult to detect with validity through non-experimental studies. Second, there has been a lack of plausible non-causal explanations and that these have been tested for and found wanting. Third, there has been the use of some type of 'natural experiment' that serves to pull apart variables that ordinarily go together or which involves risk reversal. Fourth, there has been the finding of similar effects both in varied circumstances and using different designs. Fifth, where possible, confirmatory evidence of the supposed mediating mechanism has been sustainable either through human experiments or animal models. We recognise that these are demanding criteria and, in most instances, not all five conditions can be met. Nevertheless, there are sufficient good examples of non-experimental findings giving rise to causal inferences for our support to be unequivocal – provided, and only provided, the non-experimental studies measure up to the criteria we have suggested.

The fifth criterion of an identified mediating mechanism needs further brief comments. We accept that a true cause can be identified without the mediating mechanism being known. The relationship between insulin and diabetes, and sulphonamides and puerperal sepsis would be cases in point. Nevertheless, in both instances the causal inference was in keeping with physiological knowledge and later research went on to elucidate the mechanisms. We also note that, because of the complexity of causal mechanisms in biology, there is no unequivocal end point when it can be said that the mechanism is known. Should this be at the whole body physiological level or the level of intracellular chemistry or gene expression? The simple point is that the causal inference is greatly aided by knowing something about how the cause operates.

9.3 Can non-experimental studies be misleading?

We need to turn now to the opposite question; can non-experimental studies be seriously misleading? We conclude that they can, although the published examples where this has been the case are less numerous than has sometimes been claimed. Apart from the usual problems inherent in findings from any form of research that has been poorly conducted, we picked out three main characteristics of misleading non-experimental findings. First, even with well conducted non-experimental studies using large samples, causal inferences should be very tentative if the effects are small. In no way do we doubt that true effects are often small. Rather, it is that with non-experimental studies, the ruling out of confounding effects is much more difficult in the case of small, apparently causal, effects. Second, causal inferences are especially hazardous if there are likely to be selection biases or indication biases that influence who is exposed to the putative causal experience, or ascertainment biases with respect to the detection and measurement of the relevant supposed risk or protective effect on the disease outcome. Third, causal claims should be treated with considerable caution if competing non-causal explanations have not been rigorously and thoroughly sought for and tested for using high quality discriminatory measures.

9.4 Why are there conflicting claims on causes?

The fourth basic question we sought to address is why there seem to be so many disagreements about claims that purport to have identified a cause of disease. The main answer is that most are based on poor quality, small-scale, preliminary studies, about half of which never get published. Many of the claims come from oral presentations at meetings, abstracts, discussions with journalists, or publications in journals not using high quality peer-review. Statistical significance of some difference may have been found but that has little meaning if the basic design was flawed. Moreover, statistics are the poor person's guide to validity. Of course, we are all poor people in that respect, but the real test in science lies in replication. The disagreements over causal claims are far fewer if attention is confined to replicated studies of high quality.

The requirements of good design, appropriate samples and rigorous data analysis are well appreciated by scientists but, all too frequently, they are given inadequate attention by journalists. Good guidelines on the reporting of science are available (Royal Society 2006; the Royal Institution, Social Issues Research Centre and the Royal Society 2001; Vandenbroucke *et al* 2007; and the TREND group - Des Jarlais, Lyles, Crepaz & the TREND group 2004) and we support them. There has been debate among scientists on the relative advantages and disadvantages of RCT and non-experimental studies for the identification of causes. We have not mainly focused on this debate because most likely causes of disease are not of a kind for which RCTs would be possible. Nevertheless, there are a few for which direct comparisons are possible. As it

turns out, if attention is confined to high quality replicated studies, there are relatively few well documented examples of major disagreements between non-experimental studies and RCTs with respect to the direction of effects, although there are rather more differences on the size of effects. The disagreements almost entirely concern instances when the effects found have been relatively small and when it is clear that there were likely to have been major biases (selection, indication or ascertainment) between the participants in the groups to be compared. In these circumstances, we conclude that RCTs (or RD designs if they can be applicable) are much to be preferred for the identification of a causal effect. Nevertheless, observational studies do have a complementary role in compensating for the restricted generalisation of many RCTs, for identifying rare (but serious) side-effects, and in charting long-term effects.

9.5 Do RCTs constitute the only satisfactory means of establishing causation?

RCTs have the huge advantage, when studying the effects of some interventions, of using randomisation as a means of making confounders (known and unknown, measured and unmeasured) equally likely in the two (or more) groups to be compared. No other method does that as well and, for that reason alone, almost always they should be the preferred choice for studying the effects of some intervention. It should be noted that randomisation can apply to areas or groups and not just to individuals.

Nevertheless, RCTs have several important limitations. First, they assess the effects of a new intervention and it cannot necessarily be assumed that a lack of that intervention constituted a prior cause of the initiation or cause of disease. Second, the scientific rationale will be undermined if either there is differential attrition between the groups being compared, or if the people willing to participate in the RCT differ markedly from those for whom the intervention is intended, or if the nature of the intervention makes it difficult (or impossible) to blind the participants or researchers from which group each individual is in. Third, unless specifically designed (and powered) to do so, RCTs may be less informative as to why an intervention works, or on whether it works only (or mainly) in particular subgroups. Fourth, RCTs often assess only one 'dose' of the intervention, thus making it very difficult to assess dose-response relationships. We emphasise that these limitations also apply to many non-experimental studies of interventions, so it is not that they are stronger, but rather that both have limitations. It is the fifth limitation, however, that matters most in present circumstances. That is, for many postulated risk or protective factors, RCTs are neither practical nor ethical.

9.6. Is there a statistical approach that completely deals with confounding variables?

As Appendix I brings out, there are crucially important analytic issues that need attention in all studies of causation. The question we pose here is whether there is either a single approach, or even a combination of approaches that solves all the problems. We conclude that there is not. To begin with, in a non-experimental study, only variables that have been measured can enter the analysis. All too often, the range of measured variables has been too narrow, both in respect of potential confounders and in respect of variables that could serve the role of instrumental variables. There is also the logical problem that the statistical analyses have to estimate what the effects would have been if the circumstances had been different and that requires all sorts of (often dubious) assumptions (see Rutter *et al.* 2001). Though frequently not the case, it should go without saying that the chosen method of analysis should be appropriate and

undertaken correctly. There is also the practical problem that statistical adjustment will always be unsatisfactory if there are major differences between cases and controls in the range of possible confounders.

Claims have been made that propensity scores could provide a solution in that they can effectively stratify (i.e. crudely equate) cases and controls on the demonstrated risk and protective factors for the exposure to the putative factor being investigated. The particular strength of propensity scores is that they show clearly the strata where there is very little overlap between groups and, hence, they provide pointers to the strata that it may be prudent to exclude.

The technique has been too little tested for any confident assertions to be made on the extent to which it succeeds in providing a matching that is equivalent to that obtained by RCTs. As its proponents rightly note, what it cannot do is take into account confounders that have not been measured, and it is inevitably specific to the samples being compared and influenced by how thoroughly and rigorously it is undertaken. We conclude that it constitutes a potentially useful analytic technique to consider, but we are doubtful if any statistical technique on its own could solve all problems. As Shadish and Cook (1999) put it, design must always trump data analysis; but, of course, statistics are crucial in the development and choice of designs, as well as in data analysis.

9.7 Recommendations and guidelines

Since research into the environmental causes of disease, often using non-experimental methods, is crucial for improving the nation's health, we offer five key recommendations based on our analysis of findings on the topic.

First, in view of the multiple strategic and technical issues involved in the identification of environmental causes, especially in their study through non-experimental methods, scientific expertise should be involved in all stages of policy development in relation to health issues, in the design of interventions, and in the subsequent interpretations of outcomes. This is particularly important in the case of non-experimental studies of environmental causes of disease, because of the uncertainties over causal inferences plus the very considerable public health implications. While we welcome recent efforts by the Government Social Research Unit, PolicyHub and others, we recommend that:

Government should build upon their recent efforts to integrate science into policymaking by increasing capacity building further by means of:

- **Embedding researchers into policy teams.**
- **Providing senior civil servants with scientific training.**
- **Seconding scientists to government.**
- **Building a cadre of 'evidence brokers' within government who are trained in both science and policy.**

Second, funding bodies should make clear to researchers that they attach considerable importance to well conducted non-experimental research that uses methods, such as those offered by natural experiments, that can provide good tests of putative environmental causes of disease. They should also emphasise the need to validate important findings in different settings and populations, and to undertake systematic reviews that bring together research findings using different methods. We recommend that:

The Research Base Funders' Forum should lead an initiative to reaffirm funders' support, where appropriate, for high quality non-experimental research into the environmental causes of disease, encourage studies to test previous findings in different circumstances, and undertake systematic reviews.

Third, since policymakers often have to make public health decisions, using existing research evidence, more quickly than further research can be generated and completed, it is necessary to integrate rigorous piloting into the implementation of new policies and practice. Such piloting may indicate the desirability of modifying the new policy and, occasionally, even the need for a complete rethink. Furthermore, because even well based policy changes may not bring about the expected benefits, it is crucial that the changes be introduced in a manner that allows rigorous evaluation, and that funds be provided for such evaluation. We recommend that:

The Department of Health, and other relevant government departments, should ensure that there is a greater emphasis on both pilot studies and systematic rigorous evaluations of the effects of interventions in developing and implementing health policy.

Fourth, the challenges inherent in the interpretation of high quality non-experimental and experimental evidence concerning the identification of environmental causes of disease means that researchers, their funders, and their employing institutions have a responsibility to analyse and present findings in a considered and balanced fashion. This should include considering possible alternative explanations for their findings that could modify their conclusions. Equally, scientific and medical journalists need to accept responsibility for accurate balanced reporting and interpretations. We recommend that:

The Research Base Funders' Forum should lead an initiative to foster responsibility for the accurate communication of non-experimental research. This should include consideration of whether it would be feasible to make accurate communication of results a requisite of funding.

Fifth, because research into the causes of disease and of good health is so important in people's daily lives, it is important to involve patients the lay public as active partners in the research process (Evans, Thornton & Chalmers 2007). Patients have experience that can enhance the choice of priorities in the identification of both causes of disease and in their treatment. Also, the wider participation of the public in the research process may enhance their understanding of what is involved in research and, hopefully, therefore their willingness to support research. Equally, the public needs to know how to make sense of research findings presented in the media. There is a lack of solid empirical findings on the value of such participation, but experience in many settings suggests that there are many gains (Stilgoe *et al.* 2006; Wanless 2002; Wilsdon & Willis 2004; Wilsdon *et al.* 2005). Accordingly, we recommend that:

The Departments of Health, Research Councils, and charities funding research into environmental causes of disease and interventions to prevent or treat disease should continue to involve the public and patient organisations by inviting them to participate in their expert scientific advisory committees.

It is also clear, however, that change cannot only come from above. It will be crucial to engage everyone in the issues involved in the process. Accordingly, we have provided sets of guidelines tailored to the roles of different stakeholder groups, across the continuum from undertaking research to its translation into policies and practice, in order to help them best utilise the rich pickings offered by research into the environmental causes of disease. These guidelines are set out at the start of this report.

9.8 Overall conclusion

We end, as we began, with the statement that non-experimental studies can give rise to sound causal inferences, but only if certain rather stringent conditions are met. However, we need to add that it is clear that the common multifactorial diseases all involve environmental influences on the causal processes. If progress is to be made in identifying such environmental causes, it is essential that the research into causes has a high priority and that such research should include a range of designs – both experimental and non-experimental. No one design is adequate on its own and the most progress has usually come from a thoughtful integration of findings from research using different strategies.

Appendix I: Statistics

This Appendix presents some statistical considerations relevant to causal inference from non-experimental studies, emphasizing structural and strategic issues rather than detailed techniques. In the interest of making the presentation of statistical issues succinct, we have adopted a rather didactic approach. We appreciate that for many researchers what we say will be well familiar, but for some it may introduce a few considerations that they may wish to consider further.

1 Estimation and identification

1.1 Estimation

The traditional tools of statistics (significance testing, confidence intervals, etc.) were developed to aid the interpretation of well conducted experimental studies, and centre around issues of separating 'signal' from 'noise' in the light of necessarily finite sample size. They aim to say something meaningful about the unknown probability process that generated the data analysed, taking due account of the remaining uncertainties. We may loosely term this complex of issues and methods 'statistical estimation'.

1.2 Identification

When we deal with non-experimental studies, significant new issues arise. In these cases, even if we had fully accurate knowledge of the data generating process, we would often still not be in a position to answer causal questions of interest. In order to address these we need to make additional assumptions (and hope to justify them). This is the area of 'statistical identification'. Questions concerning identifiability are often posed as if the answer were categorical. In practice, the empirical answer is not infrequently more graded, with some designs and assumptions offering only weak identification of some parameters. Identification issues form our main focus below.

2 The purpose of randomisation

The randomised controlled trial is generally taken as a 'gold standard' for the assessment of causal effects. Here are some statistical arguments pro and con:

2.1 Pro:

- Randomisation ensures 'internal validity': when we compare outcomes in different treatment groups, we are comparing like with like.
- External intervention to apply a treatment ensures that observed associations between treatment and response can be given a causal interpretation.
- Because application of treatment precedes measurement of outcome, the direction of causality is clear.
- Known probabilistic randomisation processes can be used to justify the applicability of certain statistical analyses.

2.2 Con:

- Randomisation does not ensure 'external validity': subjects entered in the experiment may not be representative of the broader population of interest.
- If we measure covariates, we may discover differences between groups that invalidate the property of 'equality of ignorance' that underlies internal validity.
- In medical studies, in particular, there will be other important but non-statistical issues, e.g. arguments for double blinding, or pragmatic or ethical obstacles to randomisation.
- Attempts to adjust for observed covariates cannot be justified by appealing to the randomisation distribution.
- There is dispute as to whether randomisation based statistical inference can be put on a firm logical foundation.

2.3 Observational studies

If, rather than experiment, we collect data passively, many of the above pros and cons

are interchanged. We can no longer guarantee internal validity. Interpreting associations causally is fraught with dangers, and in particular it can be hard to distinguish direct from reverse causation. And there is no 'objective' randomisation distribution on which inferences could be based. On the other hand, external validity may be more plausible, while appropriate adjustment for observed covariates might possibly help counter some of the inherent biases. Natural experiments can sometimes provide an alternative to the deliberately constructed randomisation distribution.

3 The problem of confounding

When we examine and compare patient responses in two or more treatment groups or levels of a risk factor in an observational setting, what we get to see is a combination of two quite distinct effects:

Treatment effect: The specific power of the treatment administered to make a difference to the outcome of interest.
Selection effect: The fact that the treatment groups are not completely random subsets of the population of interest.

Selection may be by self-selection, as when we evaluate a government initiative with voluntary participation: those individuals who choose to take part tend to be more motivated and may be, for example, receiving higher incomes. Or we may have external selection, as when a physician gives the treatment s/he prefers to those patients s/he thinks will benefit the most. Or (a hybrid case), the doctor may prescribe a treatment, but the patient self-selects whether or not to take it. In such cases, even when there is no real treatment effect whatsoever, the existence of a differential selection effect would typically lead to systematic differences between the outcomes in the different treatment groups - because we are not comparing like with like. This is the essence of the problem of confounding. It is an 'identification' issue that would not go away even if we could have perfect knowledge of the probability distributions of the response in the two treatment groups.

4 The problem of reverse causality

Reverse causality, where the process that gives rise to the outcome influences the process of selective exposure to the treatment or risk factor, is not uncommon. This is especially the case for diseases that may have an extended prodromal, or mild but chronic, early phase in which symptoms may influence behaviour, or when biomarker or other intermediate outcome data are available that may influence decision making. In economics this is referred to as endogeneity. Without isolating some exogenously determined source of variation in the treatment, even models that analyse both outcome and treatment jointly may be poorly identified. It may also be less than clear from where the identification of such models arises, and hence what are the critical assumptions.

5 Breakdown of randomisation in many trials

The trend towards taking RCT evidence as the only decisive evidence for treatment efficacy has led to their application to an ever increasing range of treatments and diseases and areas where maintenance of a rigorous trial protocol becomes more difficult. It is now not uncommon for trials to 'break down' with, for example, only a minority of patients complying with the assigned treatment regime. While the simple randomisation assigned group difference (the 'intention-to-treat' effect) may still be valuable for some public health purposes, estimating a causal effect of a treatment on patients who received no treatment is clearly problematic. However, non-compliance re-introduces the problem of selection bias: those who complied and thus received treatment may be systematically different from those who

did not. Thus the question arises: Are RCTs which break down no better than observational studies, and equally flawed by the problem of confounding? To a substantial extent the answer is 'yes, but not quite'. The 'yes' part indicates that trials should follow the practice of observational studies and extend the range of potentially confounding variables that they measure and control for. But 'the not quite' has been the focus of much recent work in the design and analysis of trials to determine how randomisation can continue to be exploited as a point of leverage for causal inference.

6 Theoretical frameworks

Statisticians have developed various formal frameworks to represent and manipulate causal relationships and infer them from data. These frameworks, while sharing many common features such as helpful algebraic and graphical methodological tools (Pearl 2000), can vary substantially in their detailed ingredients and assumptions, and especially in their (often implicit) underlying philosophies. These differences have led to some disagreements about the validity of certain purported causal inferences. However, for current purposes these are not likely to be important.

6.1 Potential responses

One popular approach (Rubin 1974; Rubin 1978) interprets a causal effect as the difference between the outcome values for a subject under two (or more) different treatments. Since it is never possible to observe more than one of these values, this approach has developed assumptions, methods and interpretations for handling such 'potential' or 'counterfactual' responses. From this point of view, confounding is understood as an association, under observational conditions, between the treatment a subject receives and the whole constellation of his potential outcomes. Causal inference then requires assumptions about the joint distribution of all these variables, together with other relevant quantities.

6.2 Decision theory

A quite distinct approach that may be more straightforward and helpful is 'decision theoretic' (Dawid 2002). This aims to estimate and compare the different probability distributions for the outcome variable, under different treatment interventions that might be applied to a subject. In an observational study, confounding holds when we have different distributions for the outcome (for given treatment) in the observational and interventional settings. Causal inference then relies on appropriate assumptions to relate the available observational distribution to the interventional distributions of real interest. In either framework, perhaps the most important contribution has been the clear identification of just what assumptions are required to support the desired causal inferences. Then ideally (if not always in practice), appropriate argument can be made for the applicability of these required assumptions in the specific case at hand.

6.3 Effect heterogeneity

Statisticians usually approach the concept of heterogeneity of effect through the concept of a random coefficient that varies from subject to subject. Standard linear model theory, and much routine experience in the non-linear (e.g. logit) case when the variation is modest, has led many researchers to expect approximate orthogonality in mean and variance of a random coefficient. Thus, many epidemiologists do not routinely expect that postulating treatment heterogeneity would much influence an estimate of the average treatment effect: commonly one might report some coefficient and leave it unspecified as to whether this is an estimate of an effect assumed common to all members of the population or an average of some population distribution of effects, and if so, over which population or sub-population.

Work in economics (Angrist *et al.* 1996; Imbens & Angrist 1994) has emphasised the importance of being much more explicit in what interpretation we are claiming for our effect estimates. Most specifically, in a context

of effect heterogeneity, we not only need to consider whether exposure to the risk factor or treatment of study might be selective with respect to confounders, but also whether it might be selective with respect to the effect distribution — for example, those who are likely to benefit most may be preferentially exposed to the treatment. This has the implication, both for observational and for most practical trials, that we need to qualify many estimates of effects as being 'local', meaning that they are relevant only for the particular sub-population that our experimental manipulation or observational setting has induced to be treated or exposed. Further distinction as to whether we are describing an average over this sub-population, or the effect at the threshold or margin that relates to those whose exposure or treatment would change as a result of minor change in the 'inducement', also becomes important.

One obvious consequence is that effect estimates are likely to vary from study to study, not necessarily as a result of different causal mechanisms, but as a result of variation in the sub-population exposed. A further consequence is that designs and analyses that may provide equivalent effect estimates when homogeneity of effect can be assumed, may no longer do so when heterogeneity is allowed. We highlight one such in paragraph 9.2 of this appendix.

7 Confounders

In describing ways of dealing with confounding, we shall here largely take the decision theoretic approach of paragraph 6.2. Similar considerations can be based on the potential response approach.

A covariate is an attribute of a subject that can (at least in principle) be measured prior to the point of treatment or exposure. Taking due account of covariates is important for three distinct reasons:

1. Covariates can affect whether or not a subject participates in a study.
2. In an observational study, a covariate can be associated with the treatment applied.
3. Covariates can affect a subject's response.

A covariate having properties two and three is called a *confounder.* In the presence of confounders we will have confounding(!), and simple statistical analyses are likely to be misleading. Even without confounding, if number one holds, care must be taken in generalising from the study to a population of interest.

7.1 Sufficient covariate

More formally, a set of covariates X is termed sufficient (for causal inference about the response Y to treatment T) if the conditional distribution of Y given X and T is the same for both the observational and the interventional setting. In this case, by making appropriate adjustment for X we can eliminate the problem of confounding. Without such adjustment - which will be unavailable when X is not measure-X is a confounder; with it, we might more appropriately call X an unconfounder.

Note that whether or not X is sufficient depends on the specific interventional and observational regimes considered, and the response variable Y. Moreover, a sufficient covariate need not be uniquely determined: for example, in a completely randomised experiment any set - even an empty set - of covariates will be sufficient.

8 Covariate adjustment

Suppose we can assume that some measured set X of covariates is sufficient. Then we can estimate the interventional distribution of Y given X and T from the observational data. Typically this will involve fitting a statistical model for the dependence of the response Y on X and T. If this can be done reliably, then we will have the information we need to decide on the treatment of a new subject S on whom we can measure X. Under certain conditions we

can apply similar methods even when X is not sufficient. For example we might consider that, because of selection (see 1 of paragraph three above), the mean of the conditional distribution of Y given X and T is greater by some amount in the observational study than its value in an interventional setting. So long as that increase can be assumed the same for all treatments considered, appropriate comparisons between the treatments, based on the observational data, will still be valid.

8.1 Difficulties

Missing information: We may not be able to measure X on the new patient S before having to choose the treatment. But if we can reasonably assume that the difference between the expected responses for the different treatments is the same for all values of X, we can just estimate this difference from the data and apply it to S. Otherwise, for each treatment we will need to take a further expectation of the conditional response distribution over the distribution of X. Sometimes this will be available from external sources (e.g. for a patient of unknown sex, we might assign equal probability for male and female). Alternatively, we might estimate this distribution from the observational data - but this could be misleading in the presence of selection effects.

Data dredging: A general statistical problem, by no means confined to causal inference, is that, when we estimate a model from limited data, there is a danger that what we find is distorted by 'random noise' in the data. This becomes more problematic as the model becomes more complex (e.g. with many variables and parameters) in relation to the number of observations: then, paradoxically, the better our chosen model appears to fit the data, the worse is it likely to perform when applied to new subjects.
Some limited protection is offered by taking account of the estimated uncertainties with which most statistical estimation techniques hedge their conclusions; but, given the many ways in which a finite set of data can be 'dredged' for apparently interesting messages, formal methods are unlikely to correct adequately for this effect.

8.2 Variable reduction

In a covariate model for Y, as above, the distribution of Y will be typically governed by some reduction V of X (which, however, depending as it does on the unknown parameters, will itself be unknown). This is true of, for example, analysis of covariance models for continuous Y and logistic linear models for binary Y, in which case V is the univariate 'linear predictor' based on X. Such a reduction will itself constitute a sufficient covariate. The model fitting process can thus be interpreted as an attempt to identify an appropriate sufficient reduction V of the original sufficient collection X. However, the possible distorting effects of 'data dredging' on this identification must always be borne in mind. Further, because an estimated reduction will always differ from the true reduction, it will not be exactly sufficient, so that transferring the estimated model from the observational to the interventional setting can introduce biases.

9 Propensity scoring

Rather than fit a model for the dependence of the response Y on X, one can fit a model for the dependence, in the observational setting, of the assigned treatment T on X. It can be shown that, if this depends on X only through the value of some reduction U of X, then U will be a sufficient covariate. For the case of binary T, a suitable choice for such U is the probability, given X, of receiving active treatment: this is termed the propensity score (Rosenbaum & Rubin 1983).

Propensity analysis proceeds via two distinct stages: first we attempt to identify the sufficient reduction U, using data on X and T only; then we estimate the dependence of Y on U and T, or condition on U through matching, or use inverse propensity scores as 'probability of treatment' weights to obtain a weighted sample in which T and U are uncorrelated (Kurth *et al.* 2006).

Since U is sufficient, average treatment effects estimated in this way this should be transferable to the interventional setting.

There are several advantages of propensity scoring over the more usual covariate adjustment. Construction of propensity scores U can (and should) be undertaken without knowledge of the outcomes Y. Consequently the same initial analysis can be used for a range of response variables. The method often also identifies extreme sub-populations of subjects who are almost certain to receive the same treatment. In such a case we will lack suitable subjects with whom to make a comparison of treated and untreated outcomes (Kurth *et al.* 2006). The propensity score approach, particularly if undertaken blind to outcome data, provides a proper basis on which such sub-populations may be excluded from the analysis.

9.1 Difficulties

In principle, exactly the same caveats about data dredging and non-exact sufficiency apply to estimating a propensity score U as to estimating the sufficient covariate V of Paragraph 8. It is often said that these are less of a concern in the propensity setting (Joffe & Rosenbaum 1999), but the evidence for this is weak.

Another concern is that the propensity score, and the performance of the associated analysis, can be highly dependent on the initial choice of sufficient covariates X.

9.2 Confusion of conditional and marginal estimators

In many of the models used for estimating homogeneous effects, although the assumptions being made may differ, the parameter being estimated is the same. Thus, effect estimates derived from, say, adjustment for covariates, or weighting by the inverse of propensity score, can be meaningfully compared, even though the former is a 'subject specific' estimate while the latter is a marginal or 'population average' estimate. However, in epidemiology the most common effect estimator is the odds ratio, commonly derived from some logistic regression. In this case the subject specific and population average estimators are targeting different parameters, which are measuring the effect in different ways — though approximate transformation from one to the other may be possible (see for example Pickles 1998). Unfortunately, papers comparing different methods of causal modelling and inference sometimes fail to distinguish differences arising from different techniques from those arising from differences in the target parameter.

10 Instrumental variables

The methods discussed so far deal only with measured confounders: a severe limitation. Much recent effort in statistical methodology has been to elaborate methods that when combined with appropriate theory and design can deliver conclusions about causality in the presence of unmeasured confounders. These all exploit, by design or assumption, some element of randomisation. That with the longest tradition is the instrumental variable (IV) approach, which attempts to identify a variable Z that, in essence, supplies a source of variation in exposure that is equivalent to randomisation, i.e. that is not correlated with variation in exposure due to confounders. Such an 'IV' Z should be strongly associated with the exposure of interest, and should not influence the outcome except through its effect on increasing or decreasing the treatment or exposure, an assumption known as the 'exclusion restriction'. In the simple case estimates can be derived using two stage least squares, where the predicted values from a first stage regression of exposure or treatment on X and Z are used as the treatment/exposure covariate in the second stage regression of Y on X and treatment/exposure.

10.1 Theory, natural experiments and Mendelian randomisation

The plausibility of regarding a variable as an instrument largely rests on experimental randomisation, or on exploitation of natural

experiments, or on theory. Direct appeal to theory has often appeared more persuasive in the social than biomedical sciences. However, a more systematic evaluation of biomedical theory may yield more generally accepted instruments. Direct testing of assumptions such as the exclusion restriction is not directly possible, although indirect evidence of the lack of association of the instrument with other confounders is often presented (Hernán & Robins 2006).

Identifying plausible IVs is becoming an important epidemiological 'game'. The prospect of genetic instruments has been especially alluring. This forms the basis of the methodology known as 'Mendelian randomisation'. A gene that is known to have a substantial influence on a risk factor of interest is used to provide a source of variation in the level of the risk factor. It is then assumed that this gene has no other direct effect on either outcome or confounders (Didelez & Sheehan 2005).

10.2 Local effect estimation and effect heterogeneity. Local IV.

When the exposure is categorical the estimates derived from IV approaches are similar or identical to those of a number of other approaches that have been developed for evaluating intervention studies with imperfect compliance. Non-compliance means that treatment receipt is selective in spite of random assignment. However, under random assignment the expected proportions of compliers and non-compliers in treated and control groups may be equated. This, together with an assumption of no effect of assignment to treatment on outcome in the absence of treatment being received, and also of monotonicity (that for all subjects the probability of receiving treatment is at least as high if they are assigned to the treatment group), yields an estimator of 'Complier Average Causal Affects' (CACE) or 'Local Average Treatment Effects' equivalent to the IV estimator (where random assignment is the IV). One formulation allows the CACE estimate, for a given setting of the X variables, to be insightfully presented as the difference in expected outcome (under the two assignments) divided by the difference in the propensity to be exposed to treatment. Estimates of this kind can be derived for several, rather than just two levels of treatment exposure, but have not yet been derived for a continuous measure of compliance. Moreover, under these circumstances, unless effect homogeneity is assumed it is not clear what the ordinary IV estimator is estimating. Instead, a local IV estimator has been proposed (Heckman & Vytlacil 1999).

11 Use of latent variable methods to test causal inferences

Particularly in mental health, much use is made of SEMs often as if they were synonymous with causal models. They are not. Many SEMs are merely models of association. SEM does, however, have several features that when combined can provide a powerful approach. Even when a confounder has been measured, if it has been measured unreliably then covariate adjustment can be imperfect. Under assumptions about the conditional independence of measurement errors, suitable data allow such a confounder to be represented in a SEM as a latent variable and for its effects to be more completely taken into account. With time varying exposure and repeated measures of outcome then a SEM with a latent variable representing a subject specific time persistent propensity for the outcome can be estimated. Allowing this to be correlated with exposure gives a class of models that provide effect estimates analogous to those described as fixed effect estimates by economists and conditional estimators by statisticians, and that give exposure outcome coefficient estimates that may be given a causal interpretation in the presence of time fixed residual confounding. Latent variables representing the effects of residual confounders can also be identified

by the inclusion of restrictions in the model, notably those that conform to the assumptions of the IV approach described in Section 10 of this appendix. However, where homogeneity in the effect of the exposure of interest does not hold, the interpretation of the effect estimates as average causal effects can rarely be assumed.

12 Objectivity and planned analysis

More sophisticated methods of analysis may help. However, RCTs differ from observational studies not just in terms of design; they also differ with respect to their whole approach to objectivity (Rubin 2007). Recognition of the need for greater objectivity does not require any questioning of the integrity of individual scientists: the factors that influence research are simply too numerous and too pervasive to keep them all in check by relying solely on every scientist's constant vigilance. Compared to observational studies, RCTs take more robust practical steps to achieve objectivity.

For example, in RCTs precise specification of the outcome measure and of the analysis that will be undertaken are made prior to obtaining the data, and overseen by a Data Monitoring Committee. In the longer run, we may find an adaptation of such an approach may serve epidemiology better than the near untestable reliance on the ability of scientists and the research process to remain unbiased throughout. Initially this might look incredibly cumbersome, seemingly ruling out the ability to deal with the many complications of data analysis, such as confounding, in any sensible and necessarily post-hoc way. However, several of the approaches to analysis we have described allow preparatory and extensive data analysis to be undertaken without the need to see the outcome data, i.e. can be undertaken blind to critical data that might bias findings. For example, this is true for the calculation of propensity scores, where extensive exploration may be necessary for their specification of what factors influence risk exposure, but which can all be undertaken without knowledge of the outcomes. Once the propensity scores have been calculated, the final analysis of the outcome is often simple and could be easily specified in advance, in much the same way that analyses are pre-specified in the Analysis Plan of an RCT, and overseen by a Data Monitoring Committee. Similar procedures to achieve objectivity may be achievable where IV are pre-specified. In the future, considerations of this kind may play as important a role in our choice of method and in our analysis strategy as the more familiar methodological considerations emphasised here.

13 Publication bias

Systematic review methodology has brought to the fore issues relating to study heterogeneity and bias. If studies have been non-selectively reported, we expect symmetry in the distribution of effect estimates across studies, and we have a good idea how the variance of that distribution should vary with sample size. Plots of published effect estimates by sample size can thus be used as evidence for publication bias. Numerous cogent examples have found an under representation of small and medium-sized studies with zero or negative effect estimates. Extensions of systematic review to identify some of these under represented studies can rarely be fully effective. Adjustment for estimated bias is possible, and should be considered as one element of a sensitivity analysis.

14 Other aspects of design and measurement important to validity

In the contrast between RCTs and observational studies the importance of randomisation has been emphasised. However, the importance of many other design elements required for observational studies should not be overlooked. The reduction of measurement errors is important, but their inevitability means that care is also required in recognising the potential

impact of correlated errors to ensure balance and, where possible, blindness to critical prior or contemporaneous data.

15 Combining experimental and non-experimental data

A design possibility that appears to have been rather little exploited is the scope for simultaneous analysis of studies of different designs, in particular the joint analysis of experimental and non-experimental studies. This would offer some scope for evaluating the relative magnitude of various biases.

16 Sensitivity analysis

Several authors have argued that statistical analysis in epidemiology should concern itself, not just with uncertainty associated with random sampling error, but also with uncertainty due to departures from the many assumptions required both for analysis and the various additional inferential steps required to generalise findings and beyond those usually considered. Sensitivity analysis could involve considering a range of different scenarios. Susser *et al.* (2006) argued that such an approach would diminish the number of fragile associations that are declared causal. Currently, the publication process offers little positive incentive to do this thoroughly. An arguably more coherent method is to represent the potential departures from assumptions by following the practice of Bayesian statistics of allowing parameters to be distributed over a range of values representing different levels of bias (Greenland 2005). While formulation of appropriate multivariate prior distributions and their identifiability poses significant problems the approach would allow a direct representation of uncertainty due to sensitivity within a generalised confidence interval.

17 Decision under uncertainty

Even the best conducted investigations rarely provide definitive answers to questions of interest. Evidence from different studies must be synthesised: some methodology exists and an infrastructure for doing this has grown up (see, for example, www.cochrane.org). That this is the case has implications for how one should evaluate proposed studies. Studies should be assessed for their potential to add and be combined with existing evidence. This puts emphasis on design and rigour rather than sample size alone. But merely seeking the best additional evidence is not always possible. It is often not an option to wait for firm answers before acting: inaction is itself an action, with consequences of its own.

Statistical Decision Theory (Raiffa 1968) supplies formal methods to guide action under uncertainty. A decision maker (DM) should first quantify all relevant consequences (combinations of available decisions and their possible outcomes) on two dimensions: the *uncertainty*, measured by *probability*, of the outcome given the decision, and the *desirability*, measured by *utility,* of the overall consequence. The best decision is that maximising expected utility. In our context this calculation needs to be undertaken with respect to the population, taking into account variations among individuals in net utility and alternative options with respect to the targeting or restriction of an exposure or intervention.

Regarding desirability, it is obvious that this will always involve an irreducibly subjective element. Formal methods exist to help DM construct his or her utility function, for example for multi-attribute consequences; but these can not bring differing opinions into line. Nor can one say much, in general terms, to help address problems involving multiple stakeholders. The existence of groups with contrasting valuations of benefits could be integrated into the decision making process along the lines similar to those for differing

assessments of prior evidence (Spiegelhalter *et al.* 1994), but much remains largely informal.

The situation with regard to uncertainty would appear rosier — if only we could get agreement over probabilities. But this is not unproblematic.

It is helpful to distinguish a variety of types of uncertainty, differing in particular in their degree of 'objectivity':

Objective chance: Even with full knowledge of the relevant processes, the future remains unpredictable. Probabilities based on such 'full knowledge' can be considered 'objective'. However, such knowledge is typically unattainable except through massive experiments.

Parameter uncertainty: We might know only the general structure of the underlying process, its full features being determined by a currently unknown 'parameter'. If we can obtain data from the same process we can learn about its parameter — though always imperfectly. As described in Section one, quantification of the residual parameter uncertainty is the principal task of 'statistical estimation' theory. And often this will be reasonably objective, in the sense that divergent initial opinions will be brought into essential agreement by moderately sized experiments.

Model uncertainty: Not knowing even the process structure, we could entertain a variety of possibilities. Again, data from the target process can help to choose between these models - but the residual model uncertainty is often much more sensitive to initial opinions. Evidence synthesis: We will often have relevant evidence, perhaps targeted at a variety of parameters, from a range of different sources and studies. Methods for combining all the evidence typically involve strong subjective assumptions and inputs. Only when there is broad agreement that the assumptions are reasonable, and there is extensive and broadly consonant evidence, is a clear conclusion likely.

The only formal statistical methodology that supports the quantification of all the above kinds of uncertainty — explicitly accounting for their subjective elements, and supporting their integration into Statistical Decision Theory — is supplied by the theory of Bayesian Inference (Bernardo & Smith 1994). In particular this describes how uncertainty (measured by probability) should be coherently updated in the light of new evidence.

While the above considerations apply equally to non-causal and causal inferences, for putative causal inferences from non-experimental data there is the additional problem that any conclusions are likely to be very sensitive to non-testable assumptions about the relationship between the processes generating the observed data and those governing the desired inference. For example, there may be disagreement as to what constitutes a set of sufficient covariates, or what is the appropriate way to adjust for them. Such disagreements can often render any inferences made highly subjective. Although sometimes considered as another argument favouring experimental over non-experimental studies, similar issues arise when experiments are considered in context. Thus trial data may be analysed from the point of view of enthusiasts and sceptics, the opinions of each being represented by some prior distribution of likely effect size, and it then being necessary for the trial evidence for treatment effect to convince a sceptic before a case for early stopping of the trial can be made (Spiegelhalter *et al.* 1994). Such a representation of diversity in the decision making process need not be limited to the assessment of prior evidence nor to the RCT context.

Whenever a decision needs to be taken, this should be done in the clear and full light of all current uncertainties, which should be made explicit and justified to the greatest extent possible. It is also valuable to explore sensitivity of the 'optimal' decision to changes in assumptions (Ades *et al.* 2006).

Appendix II: Working group and summary of their interests

The report was prepared by an Academy of Medical Sciences working group. Members participated in their own capacity rather than on behalf of their affiliated organisation.

Professor Sir Michael Rutter CBE FRS FBA FMedSci, Vice-President, Academy of Medical Sciences (chair)
Sir Michael has engaged in much research to identify possible environmental causes of disease, including the use of natural experiments. He has received multiple grants and contracts from government departments and has served on various governmental advisory groups, including that concerned with SureStart. He has given evidence to various hearings and advisory groups on putative causes of disease, including lead, intrauterine alcohol exposure and vaccines. He has served on grant-giving committees for both Research Councils and charities and has been active in public engagement with research.

Professor Philip Dawid, Professor of Statistics, University of Cambridge
Professor Dawid has a long-standing academic interest in the philosophical and methodological issues of making causal inferences from statistical data. He is an advocate of the 'decision-theoretic' probabilistic approach, which is somewhat at odds with more popular counterfactual formulations. He has been a member of the UK Medicines Commission, and has received substantial funding from the Leverhulme Trust and the Economic and Social Research Council (ESRC) for an interdisciplinary research programme on 'Evidence, Inference and Enquiry'.

Dr Aroon Hingorani, Reader and Honorary Consultant and British Heart Foundation Senior Fellow, University College London
Dr Aroon Hingorani is supported by a Senior Research Fellowship from the British Heart Foundation (BHF) and research awards from BHF and MRC. He serves as a member of the Editorial Board of the Drug and Therapeutics Bulletin (BMJ publishing group) and sits on the Scientific Advisory Board of London Genetics. He is a current member of the National Institute of Public Health and Clinical Excellence Guideline Development Group on Prevention of Venous Thromboembolism.

Dr Richard Horton FRCP FMedSci, Editor, The Lancet
Dr Horton edits a medical journal that publishes experimental and non-experimental research. He has a strong interest in clinical epidemiology and its relation to publication and research ethics. He has also engaged in research to understand the way science is written and interpreted. He was a founder member of the UK Committee on Publication Ethics and has spoken and written widely on controversies in the health sciences. He has taken part in several inquiries, working parties, and committees on research integrity.

Professor Peter Jones FMedSci, Professor of Psychiatry, University of Cambridge
Professor Jones undertakes epidemiological research using observational and randomised designs to investigate causes and treatments of mental illness, with an emphasis on life course approaches and the interaction between environmental and personal characteristics. His research is supported by grants and contracts from Research Councils, the NHS, charities and industry; he has served on grant-giving panels in each of these contexts. He has contributed advice to policy and strategy committees in the public sector including those considering disease causation, such as the reconsideration by the Advisory Council on the Misuse of Drugs of the classification of cannabis in the light of associations with psychotic disorders.

Professor Kay-Tee Khaw CBE FMedSci, Professor of Clinical Gerantology University of Cambridge
Professor Khaw aims to identify what we can do to improve health and prevent chronic diseases including cardiovascular disease, cancer and osteoporosis in the population through better understanding of the biologic, behavioural and wider socioeconomic determinants of health. Her research programme is based on large population observational and interventional studies.

Dr Bill Kirkup, Director General for Programmes, Department of Health
Dr Kirkup, formerly a practising clinician and public health doctor in the NHS, has been a senior civil servant in the Department of Health since 1996. Since then he has advised on many aspects of both health and NHS policy. His own former research interests have necessarily taken second place, but he has published some epidemiological studies, for example on trigger factors for heart disease and stroke, and some observational papers on health care in conflict-affected countries.

Dr Geoff Mulgan, Director, Young Foundation
Dr Mulgan has been involved in research in various roles, both in and out of government. As director of the think-tank Demos, he oversaw research projects funded by several national governments, the ESRC, businesses and foundations. As head of the Government Strategy Unit he had oversight of the Chief Social Researcher and her team, and commissioned, both directly and indirectly, many research projects. As a member and later head of the Downing Street Policy Unit he was also a significant user of research. He is currently director of the Young Foundation which primarily focuses on practical initiatives but also undertakes research, including some funded by the European Commission, the ESRC and a wide range of other foundations. He is a visiting professor at several universities - LSE, UCL and Melbourne. He sits on the board of the Work Foundation and the Design Council and is chair of Involve and of the Carnegie Inquiry into the future of civil society.

Professor Catherine Peckham CBE FMedSci, Professor of Paediatric Epidemiology, Institute for Child Health
Professor Peckham has engaged in non-experimental research in relation to the national birth cohort studies and observational studies, particularly those relating to infections in pregnancy and childhood. Her most recent research has focused on perinatal HIV infection. Her work has been supported by grants from the MRC, Wellcome Trust and other charities, the European Union and the Department of Health. She has served on government advisory committees, grant-giving boards and international bodies in science and health. She has been deputy chair of the Nuffield Council on Bioethics, a member of the Advertising Standards Authority and has an interest in the public understanding of science.

Professor Andrew Pickles, Professor of Epidemiological and Social Statistics, University of Manchester
Professor Pickles has participated in research across a wide range of social, psychological, and bio-medical topics in the UK and US. Current research and teaching interests include causal analysis in life-course epidemiology and the formulation, estimation and software for multilevel SEM. His research has been funded through multiple grants from Research Councils and charitable trusts. He has served on various advisory groups, primarily concerning longitudinal studies.

Professor Robert Souhami CBE FMedSci, Emeritus Professor of Medicine, University College London
Professor Souhami has for many years conducted therapeutic research in cancer, including large scale clinical trials. He has served on grant awarding committees in the field of clinical cancer research and was the Director of Clinical Research at Cancer Research UK and subsequently Executive Director of Policy and Communication for that charity. He serves on scientific advisory boards for cancer centres in UK and in France and has been chair of academic working parties concerning aspects of medical research, especially with respect to cancer.

Dr Geoff Watts FMedSci, Freelance Science and Medical Journalist
Dr Watts is a freelance science and medical journalist who broadcasts on these matters for BBC Radio, and writes about them for a variety of lay and professional publications. He also acts as a paid or unpaid consultant to many governmental, charitable and commercial organisations, including drug companies. This work is principally concerned with issues of public engagement, and with the communication of science and medicine both to lay people and to professionals.

Laurie Smith, Senior Officer, Medical Science Policy, Academy of Medical Sciences (Secretariat).

Jenny Wickham, PA to Professor Sir Michael Rutter, Institute of Psychiatry, Kings College London (Secretariat).

Appendix III: Reviewers

The review group was appointed by the Academy's Council:

Professor John Savill FRSE FMedSci (Chair)
Vice-Principal and Head of the College of Medicine and Veterinary Medicine, University of Edinburgh

Professor Yvonne Carter OBE FMedSci
Dean and Pro-Vice Chancellor (Regional Engagement) Warwick Medical School and the University of Warwick

Professor Rudolph Klein CBE FBA FMedSci
Visiting Professor, London School of Economics

Professor Sally Macintyre CBE FRSE FMedSci
Director of the Social and Public Health Sciences Unit, MRC

Professor James Robins
Mitchell L and Robin LaFoley Dong Professor of Epidemiology, Harvard University

The Academy is also grateful to the following people for providing comment on the draft report:

Professor David Fergusson FRS NZ
Executive Director, Christchurch Health and Development Study, Christchurch School of Medicine and Health Sciences, New Zealand

Professor Klim McPherson FMedSci
Visiting Professor of Public Health Epidemiology, University of Oxford

Professor Terrie Moffitt FBA FMedSci
Professor of Social Behaviour and Development, Institute of Psychiatry

Professor Ezra Susser
Anna Cheskis Gelman and Murray Charles Gelman Professor and Chair, Columbia University, USA

Mrs Hazel Thornton Hon D.Sc. (Leicester)
Honorary Visiting Fellow, University of Leicester

Professor Nicholas Wald FRS FMedSci
Director of the Wolfson Institute for Preventative Medicine, Barts and the London School of Medicine and Dentistry

Appendix IV: List of consultees and respondents to the call for evidence.

Responses to the call for evidence

Organisations

Arthritis Research Council
Association for Spina Bifida and Hydrocephalus
AstraZeneca
British Academy
Cochrane Collaboration
Department of Health
Economic and Social Research Council
Health Protection Agency
Medical Research Council
Mental Health Foundation
National Cancer Research Institute
National Institute for Health and Clinical Excellence
Royal College of Anaesthetists
Royal College of Paediatrics and Child Health
Royal College of Physicians
Royal College of Psychiatrists
Royal College of Radiologists
Social Issues Research Centre
Stroke Association
UK Clinical Research Collaboration
Wellcome Trust

Individuals

Professor Nancy Adler, University of California, San Francisco, USA
Professor Nick Black, London School of Hygiene and Tropical Medicine
Sir Walter Bodmer FRS HonFRSE FMedSci, University of Oxford
Professor Norman Breslow, University of Washington, USA
Dr Annie Britton, University College London
Professor Nancy Cartwright FBA, London School of Economics
Sir Iain Chalmers FMedSci, James Lind Institute
Professor David Clayton, University of Cambridge
Professor David Coggon OBE FMedSci, University of Southampton
Professor Rory Collins FMedSci, University of Oxford
Professor Marcelas Contreras FMedSci, National Blood Service
Professor George Davey-Smith FMedSci, University of Bristol
Dr Hiten Dodhia, Lambeth Primary Care Trust
Ailsa Donnelly, Royal College of General Practitioners Patient Partnership Group
Professor Jeffrey Drazen, New England Journal of Medicine
Professor George Ebers FMedSci, University of Oxford
Professor Martin Eccles FMedSci, Newcastle University
Professor Griffith Edwards CBE FMedSci, Institute of Psychiatry
Professor Chris Frith FRS FMedSci, Institute of Neurology

Professor Charles Galasko FMedSci, University of Manchester
Dr Ruth Gilbert, Institute of Child Health
Professor Sander Greenland, University of California Los Angeles, USA
Dr Jane Gregory, University College London
Dr Trish Groves, British Medical Journal
Professor Terry Hamblin FMedSci, University of Southampton
Professor James Heckman, University of Chicago, USA
Professor Jenny Hewison, University of Leeds
Professor Ieuan Hughes FMedSci, University of Cambridge
Professor Kenneth Kendler, Virginia Commonwealth University, USA
Professor Diana Kuh, University College London
Professor Michael Langman FMedSci, University of Birmingham
Professor Sir Michael Marmot FMedSci, University College London
Professor Klim McPherson FMedSci, University of Oxford
Professor Thomas Meade FRS FMedSci, London School of Hygiene and Tropical Medicine
Professor Jonathan Meakins, University of Oxford
Professor Terrie Moffitt FBA FMedSci, Institute of Psychiatry/Duke University
Dr Bridgitte Nerhlich, University of Nottingham
Vivienne Parry, Freelance writer and broadcaster
Professor Mike Pringle CBE FMedSci, University of Nottingham
Dr Barnaby Reeves, University of Bristol
Professor Martin Roland CBE FMedSci, University of Manchester
Professor Paul Rosenbaum, University of Pennsylvania, USA
Professor Kenneth Rothman, Boston University, USA
Dr Sharon Schwartz, Columbia University, USA
Professor Anthony Seaton FMedSci, University of Aberdeen
Professor Stephen Senn, University of Glasgow
Professor William Shadish, University of California, USA
Dr Nuala Sheehan, University of Leicester
Professor Sarah Stewart-Brown, University of Warwick
Professor Stephen Suomi, National Institutes of Health, USA
Professor Ezra Susser, Columbia University, USA
The Lord Turnberg of Cheadle FMedSci
Professor Martin Vessey FRS FMedSci, University of Oxford
Professor Nicholas Wald FRS FMedSci, St Barts & the London School of Medicine and Dentistry
Professor Graham Watt FMedSci, University of Glasgow
Dr William Whitehouse, University of Nottingham

Workshop attendees

Listed below are the names of those who attended the workshop. A summary of the workshop is available from: http://www.acmedsci.ac.uk/p47prid50.html

Susan Barber, Cancer Research UK
Sally Brearley, Patient's Forum
Tracey Brown, Sense About Science
Professor Nancy Cartwright FBA, London School of Economics
Dr Lee-Ann Coleman, British Library

Dr Peter Coleman, Stroke Association
Professor Rory Collins FMedSci, University of Oxford
Professor Sally Davies FMedSci, Department of Health
Simon Denegri, Association of Medical Research Charities
Professor Adrian Dixon FMedSci, Royal College of Radiologists
Professor Pat Doyle, The London School of Hygiene and Tropical Medicine
Dr Alan Doyle, Wellcome Trust
Fiona Fox, Science Media Centre
Dr Trish Groves, British Medical Journal
Professor Graham Hart, University College London
Professor Dave Leon, London School of Hygiene and Tropical Medicine
Mary Manning, Academy of Medical Sciences
Dr Peter Marsh, Social Issues Research Council
Professor Klim McPherson FMedSci, University of Oxford
Professor Jonathan Meakins, University of Oxford
Dr Liz Miller, Health Protection Agency
Dr Helen Munn, Academy of Medical Sciences
Katrina Nevin-Ridley, Wellcome Trust
Sir Michael Rawlins FMedSci, National Institute for Clinical and Public Health Excellence
Professor Barnaby Reeves, University of Bristol
Dr Andrew Russell, Association for Spina Bifida and Hydrocephalus
Laurie Smith, Academy of Medical Sciences
Roger Steel, INVOLVE
Dr Hazel Thornton, University of Leicester
Professor Peter Weissberg, British Heart Foundation

Appendix V: Glossary

This glossary is intended to help readers understand some of the terms used in this report; it is not presented as a definitive list of terms.

Absolute risk	The absolute likelihood (risk) that a given outcome will occur in a person exposed to some causal agent.
Acetylaldehyde	A colourless volatile liquid used in the manufacture of acetic acid, perfumes and flavours; it is an intermediate in the metabolisation of alcohol; it is also found in tobacco smoke.
Addison's disease	A rare chronic disorder in which the adrenal gland does not produce sufficient steroid hormones.
Allocation (indication) bias	A bias created by systematic differences between the characteristics of those allocated, and those not allocated, to a particular group during an investigation.
Anencephaly	A congenital defective development of the brain; infants with anencephaly are born without a forebrain (front part of the brain) and a cerebrum (thinking and coordinating part of the brain).
Apoptosis	A normal genetically directed process of cell self destruction.
Arthroscopy	A surgical procedure in which examination and sometimes treatment of damage to a joint is performed using an arthroscope, inserted into the joint through a small incision.
Ascertainment bias	A systematic distortion in measuring the true frequency of a phenomenon due to systematic differences between cases and controls in the likelihood of the outcome in question being detected.
Atherosclerosis	A disease affecting arterial blood vessels commonly referred to as a hardening or 'furring' of the arteries.
Attributable risk	The overall effect of an identified causal element on the incidence of the disease in the general population.
Blinding (in reference to clinical trials)	A means of avoiding bias in reporting by ensuring that either patients or researchers remain unaware of which group they are in. Double blinding means that both are unaware. Blinding involves concealing intervention assignments from patients and/or the investigators.
Calcium antagonist (blocker)	Also known as calcium channel blockers, calcium antagonists prevent calcium from entering cells of the heart and blood vessel walls; this leads to relaxation of the blood vessels and a consequent decrease in blood pressure and pulse rate.
Case-control	Studies that retrospectively compare a group of patients who have a medical condition with those who do not, in order to identify factors that may contribute to the cause of the condition.
Catecholamines	A group of chemical compounds derived from the amino acid tyrosine, two of which, dopamine and noradrenaline, act as neurotransmitters (chemical messengers) in the central nervous system and hormones in the blood.
Causal graphs	Graphs that portray the causal model that is being proposed.
Cohort	A group of subjects sharing the same statistical or demographic characteristic that are followed over time as a group.

Comorbidity	The co-occurrence of two or more diseases or disorders.
Compliance	A patient's adherence to the recommendations of a healthcare professional, particularly in relation to medication.
COMT gene	A gene that controls the functioning of the catechol-O-methyltransferase enzyme.
Confidence interval	The interval of measurement within which a particular percentage (usually 95%) of scores will lie. The measure provides a useful index of the degree of precision that can be attributed to a particular score or rating.
Confounding	Where an observed association can be explained by some third variable that influences both the proposed causes and its supposed consequences.
Counterfactual	Something that could have happened, but not simultaneously, with respect to the exposure (or non-exposure) to the supposed causal influence.
Covariate	An attribute of a subject that can, at least in principle, be measured prior to the point of treatment or exposure.
Cross-sectional	A study in which a group of subjects are compared on one or more variables at a single point in time. The alternative is a longitudinal study.
Diagnostic specificity	The situation in which a causal agent has an effect on outcome that is specific to some disease or disorder.
Diethylstilbestrol	An early synthetic form of the hormone oestrogen.
Differential post-assignment attrition	A situation in a clinical trial in which there is differential attrition of the treatment and control groups after subjects have been assigned to either status.
Discordant twin pairs	Twin pairs in which the two twins are discordant for some feature.
Disinhibited attachment disorders	Disorders that involve a failure (or relative failure) in the development of selective social relationships.
Double blinding (in reference to clinical trials)	Trials where treatment assignments are concealed from both patients and investigators, in order to prevent bias. (See blinding)
Dose-response (biological gradient)	The relationship between the level of exposure (dose) of any causal agent and the effect it has on a subject.
Ecological designs	Research designs that study populations rather than individuals.
Encephalocele	A congenital neural tube defect; babies are born with a hole in the skull, through which the brain protrudes.
Endogeneity	A change that comes from a feature within.
Experimental	A set of actions or observations performed in order to verify or falsify a hypothesis (theory) or to examine relationships among variables manipulated or observed.
External validity	The degree to which the results of a research study are generalisable to other populations and circumstances outside of the sample investigated.
Fenfluramine	A drug that depresses the central nervous system regulating mood, leading to a feeling of fullness and loss of appetite.

Fibrinogen	A protein produced by the liver that is essential to the formation of blood clots.
Helicobacter pylori	A spiral shaped bacterium that lives in the stomach and the section of intestine just below the stomach.
(Familial) Hyper-cholesterolaemia	Elevated levels of cholesterol in the blood. Familial hypercholesterolemia is a rare inherited genetic disorder characterised by highly elevated LDL cholesterol and cardiovascular disease early in life.
Instrumental variable	A variable that influences an outcome through an effect on the causal agent being investigated.
Interactive term	A statistical component that reflects interactions among variables.
Internal validity	The degree to which a study produces true (valid) findings within the sample investigated.
In utero	Literally: Within the uterus.
Itraconazole	An antibiotic used for treating serious fungal infections.
Ketoacidosis	A serious medical condition in which reduced or abnormal metabolism of carbohydrates leads to the body using fat as an energy source, which results in the accumulation of toxic chemicals called ketones in the blood.
Latent construct	The underlying trait or feature that is supposed to be indexed by the specific measures used in a study.
Lipoproteins	Particles that transport fats and cholesterol through the bloodstream.
Longitudinal	A study in which the same subjects are observed over a period of time. The alternative is a cross-sectional study.
MAOA	The MAOA gene is involved in the production of the enzyme monoamine oxidase. This enzyme breaks down chemicals (neurotransmitters) that control mood, aggression and pleasure.
Mediating mechanism	The intermediate mechanism that underlies a relationship between two variables.
Mendelian randomisation	A technique that capitalises on Mendel's second law, which states that inheritance of one trait is independent of another. It is used as a technique to obtain randomisation with respect to some environmental factor thought to be causal for some disease.
Mesothelioma	A disease in which tumours or cancerous cells develop in the lining of the chest cavity, abdominal cavity or pericardium (layer around the heart).
Meta-analysis	A statistical technique used to combine several studies.
Microphthalmia	To have abnormally small eyes.
Multivariate twin design	The analysis of twin data in which two or more variables are simultaneously considered together.
Myxoedema	A disease caused by decreased activity of the thyroid gland.
Natural experiment	An experiment that utilises naturally occurring differences in observable phenomena.
Nesting	The inclusion of one study design within another.
Neural tube	The precursor to the central nervous system in the developing embryo.

Neuroendocrine	The hypothalamic-pituitary-adrenal system involved in the production of hormones that play a major role in stress responses.
Nifedipine	A drug used to treat diseases such as hypertension and angina pectoris.
Non-experimental	The systematic, often quantitative, observation of biomedical phenomena in a population without deliberately planned scientific manipulation (or control) of the variables under investigation.
Observational study	For the purposes of this report the term 'observational study' is interchangeable with the term 'non-experimental study'.
Odds Ratio	A measure of effect size.
Pellagra	A vitamin deficiency caused by a lack of niacin (vitamin B3) and protein.
Perceptual deafness	Deafness due to a nerve conduction failure rather than a middle ear obstruction.
Perimenopausal	Around the time of menopause.
Perinatal	The period after 22 weeks of gestation and seven days after birth.
Phenylketonuria	A genetic disorder characterised by a deficiency in the enzyme phenylalanine hydroxylase.
Phocomelia	A congenital disorder where the limbs are very short or absent, sometimes with flipper-like hands and/or feet.
Prodromal	A term used to describe an early phase of a disease.
Propensity score	The conditional probability of an exposure to a particular experience thought to cause disease (based on background variables).
Publication bias	The tendency for positive findings to be more likely to be published than negative findings.
Puerperal sepsis	A serious form of septicaemia contracted by women shortly after child birth or abortion.
Quasi-experiment	A research method that lacks random assignment, but which involves a variety of design features that aim to provide some equivalent to experimental control.
Random error	Findings that have arisen by chance.
Randomised controlled trial	A research design in which subjects are randomly allocated to case or control status. Randomised controlled trials are often blinded or double blinded.
Rating bias	Bias caused when prior knowledge or expectation influences reporting.
Regression discontinuity design	A research method in which allocation to case or control status is determined by the assignment variable using a strictly determined cut-off rather than randomisation.
Relative risk	Whether the risk after exposure to some causal factor is greater or lesser than that in the general population.
Residual confounders	Those confounders that remain after others have been adjusted for.
Retinopathy	Non-inflammatory damage to the retina.
Reverse causation	Circumstances in which the process that gives rise to the outcome influences the process of selective exposure to the treatment or risk factor.

Risk factors	Factors that are statistically associated with some disease or disorder outcome.
Risk ratio	See relative risk.
Sample	A subset of a population who are subject to an investigation.
Selection bias	Systematic error that arises from the way in which subjects are included in a study.
Sensitivity analysis	Statistical techniques to determine the overall strength of effects. They have been particularly employed to quantify how strong a confounder would have to be to overturn a causal inference from a case-control comparison.
Seronegative	Literally 'absent from the blood'.
Seroprevalence	The number of people in a given population whose blood test is positive for a particular factor.
Somatic disease	Disease involving bodily dysfunction or damage.
Significance level	The point at which the null hypothesis is rejected and the alternative hypothesis is accepted in a test of statistical significance. Usually five percent.
Singleton	Someone who is not a twin (or part of a large multiple birth set).
Statistical regression	A generic term for all methods attempting to fit a model to observed data in order to quantify the relationship between two groups of variables.
Statistical significance	The point at which a given association is judged not to be due to the play of chance.
Stevens-Johnson syndrome	A rare but severe condition that results in hypersensitivity of the skin and mucous membranes.
Stratify	In the context of medical research this refers to the division of a sample into groups with one or more factors in common.
Structural equation modelling	Statistical approaches that use correlations among variables to estimate some latent construct that is relevant either with respect to measurement or studying possible causal pathways.
Sulphonamide	A group of antibacterial drugs.
SureStart	An initiative by the UK Government intended to give children a better start in life through improved education, health and family support with emphasis of community development and outreach.
Systematic error	Non-random error that gives rise to bias.
Teratogenic	A factor that causes congenital malformations or gross deformity.
Thimerosal	An organic mercury compound used as an antiseptic or antifungal agent, and particularly as a vaccine preservative.
Thyroxine	A hormone involved in control of the metabolism that is secreted by the thyroid gland.
Treatment diffusion	An effect by which a treatment spreads beyond the intended target population; typically spreading to a control group.
Treatment dilution	An effect by which a treatment has less than the intended impact on the target population.

Tuberous Sclerosis	A rare genetic disorder that causes tumours to grow in the brain, kidneys, heart, eyes and skin.
Vaginal clear cell carcinoma	A rare type of cancer of the vagina that occurs in young women whose mothers took diethylstilbestrol.
Zidovudine	The first antiviral drug approved for the treatment of HIV.

Appendix VI: Abbreviations

BHF	British Heart Foundation
BSE	Bovine Spongiform Encephalopathy
CACE	Complier Average Causal Effects
CNS	Central Nervous System
DALY	Disability Adjusted Life Year
DM	Decision Maker
ECS	European Collaborative Study
ESRC	Economic and Social Research Council
G x E	Gene-Environment Interaction
HCD	Human Capital Development
HRT	Hormone Replacement Therapy
INUS	Insufficient but Necessary components of Unnecessary but Sufficient Causes
IPTW	Inverse Probability of Treatment Weighting
IV	Instrumental Variable
LDL	Low Density Lipoprotein
LFA	Labour Force Attachment
LIV	Local Instrumental Variable
MMR	Measles, Mumps and Rubella (vaccination)
MRC	Medical Research Council
MRI	Magnetic Resonance Imaging
NICE	National Institute for Health and Clinical Excellence
NTD	Neural Tube Defect
RCT	Randomised Controlled Trial
SEM	Structural Equation Models
SES	Socio-economic Status
SIDS	Sudden Infant Death Syndrome
TNF	Tumour Necrosis Factor
vCJD	New variant Creutzfeld-Jacob Disease

Appendix VII: References

Academy of Medical Sciences (2004). *Calling Time: the Nation's drinking as a major health issue*. http://www.acmedsci.ac.uk/p99puid20.html

Ades A, Claxton K & Sculper M (2006). *Evidence synthesis, parameter correlation and probabilistic sensitivity analysis*. Health Economics. **15**, 373–381.

Adler NE & Rehkopf D (in press). *U.S. disparities in health: descriptions, understanding causes and identifying mechanisms*. Annual Review of Public Health.

Anderson KE, Lytton H & Romney DM (1986). *Mothers' interactions with normal and conduct-disordered boys: Who affects whom*? Developmental Psychology. **22**, 604-609.

Angrist J, Imbens G & Rubin D (1996). *Identification of causal effects using instrumental variables.* Journal of the American Statistical Association. **91**, 444-455.

Arseneault L, Cannon M, Witton J & Murray R (2004). *Causal association between cannabis and psychosis: examination of the evidence*. British Journal of Psychiatry. **184**, 110-117.

Atladóttir HO, Parner ET, Schendel D, Dalsgaard S, Thomsen PH & Thorsen P (2007). *Time trends in reported diagnoses of childhood neuropsychiatric disorders*. Archives of Pediatric and Adolescent Medicine. **161**, 193-198.

Auvert B, Taljaard D, Lagarde E, Sobngwi-Tambekou J, Sitta R & Puren A (2005). *Randomized, controlled intervention trial of male circumcision for reduction of HIV infection risk: the ANRS 1265 Trial*. PLoS Medicine. **2**, 1112-1122.

Avorn J (2006). *Dangerous deception – Hiding the evidence of adverse drug effects*. New England Journal of Medicine. **355**, 2169-2173.

Bailey RC, Moses S, Parker CB, Agot K, Maclean I, Krieger JN, Williams CFM, Campbell RT, & Ndinya-Achola J (2007). *Male circumcision for HIV prevention in young men in Kisumu, Kenya: A randomized controlled trial*. Lancet. **369**, 643-656.

Bar-Oz B, Moretti ME, Mareels G, Van Tittelboom T & Koren G (1999). *Reporting bias in retrospective ascertainment of drug-induced embryopathy*. Lancet. **354**, 1700-1701.

Barker S (2002). *What is Addison's Disease?* http://www.addisons.org.uk/info/addisons/page1.html

Beal S (1988). *Sleeping position and SIDS*. Lancet. **2**, 512.

Bech BH, Obel C, Henriksen TB & Olsen J (2007). *Effect of reducing caffeine intake on birth weight and length of gestation: randomised controlled trial*. British Medical Journal. **334**, 409-414.

Becker HC, Diaz-Granados JL & Randall CL (1996). *Teratogenic actions of ethanol in the mouse: A minireview*. Pharmacology Biochemistry and Behavior. **55**, 501-513.

Bell RQ (1968). *A reinterpretation of the direction of effects in studies of socialization*. Psychological Review, **75**, 81-95.

Bell RQ & Harper LV (1977). *Child effects on adults.* Erlbaum. Hillsdale, NJ.

Benson K & Hartz AJ (2000). *A comparison of observational studies and randomized controlled trials*. New England Journal of Medicine. **342**, 1878-1886.

Beral V, Banks E & Reeves G (2002). *Evidence from randomised trials on the long-term effects of hormone replacement therapy*. Lancet. **360**, 942-944.

Bernardo JM & Smith AFM (1994). *Bayesian Theory*. John Wiley & Sons. Chichester, UK.

Bjelakovic G, Nikolova D, Simonetti RG & Gluud C (2004). *Antioxidant supplement for prevention of gastrointestinal cancer: a systematic review and meta-analysis*. Lancet. **364**, 1219-1228.

Bjelakovic G, Nikolova D, Gluud LL, Simonetti RG & Gluud C (2007). *Mortality in randomised trials of antioxidant supplements for primary and secondary prevention*. Journal of the American Medical Association. **297**, 842-856.

Blood Pressure Lowering Treatment Trialists' Collaboration (2000). *Effects of ACE inhibitors, calcium antagonists, and other blood-pressure-lowering drugs: results of prospectively designed overviews of randomized trials*. Lancet. **355** 1955-1964.

Bollen KA (1989). *Structural equations with latent variables.* Wiley. New York.

Borge AIH, Rutter M, Côté S & Tremblay RE (2004). *Early child care and physical aggression: differentiating social selection and social causation*. Journal of Child Psychology and Psychiatry. **45**, 367-376.

Bowlby J (1951). *Maternal care and mental health.* World Health Organization. Geneva.

Brennan P, Hsu CC, Moullan N, Szeszenia-Dabrowska N, Lissowska J, Zaridze D, Rudnai P, Fabianova E, Mates D, Bencko V, Foretova L, Janout V, Gemignani F, Chabrier A, Hall J, Hung RJ, Boffetta P & Canzian F (2005). *Effect of cruciferous vegetables on lung cancer in patients stratified by genetic status: A Mendelian randomization approach*. Lancet. **366**, 1558-1560.

Bresnahan M, Begg MD, Brown A, Schaefer C, Sohler N, Insel B, Vella L & Susser E (2007). *Race and risk of schizophrenia in a US birth cohort: another example of health disparity?* International Journal of Epidemiology, Apr 17 E-pub ahead of print.

Broman SH, Nichols PL & Kennedy WA (1975). *Preschool IQ: Prenatal and early developmental correlates*. Erlbaum. Hillsdale, NJ.

Brown R (2001). *Evidence-based policy making or policy-based evidence: The case of quality assurance in Higher Education*. Professorial lecture, City University, London, 11th December 2001.

Cahan S & Cohen N (1989). *Age versus schooling effects on intelligence development*. Child Development. **60**, 1239-1249.

Campbell DT & Stanley JC (1963). *Experimental and quasi-experimental designs for research*. Houghton Mifflin Company. Boston.

Cartwright N (2007). *Are RCTs the Gold Standard?* Biosciences. **2**, 11-20.

Casas JP, Bautista LE, Smeeth L, Sharma P & Hingorani AD (2005). *Homocysteine and stroke: Evidence on a causal link from Mendelian randomization*. Lancet. **365**, 222-232.

Case A, Lubotsky D & Paxson C (2002). *Economic status and health in childhood: The origins of the gradient*. American Economic Review. **92**, 1308-1334.

Caspi A, McClay J, Moffitt TE, Mill J, Marin J, Craig IW, Taylor A & Poulton R (2002). *Role of genotype in the cycle of violence in maltreated children*. Science. **297**, 851-854.

Caspi A & Moffitt TE (1991). *Individual differences are accentuated during periods of social change: The sample case of girls at puberty*. Journal of Personality and Social Psychology. **61**, 157-168.

Caspi A & Moffitt TE (2006). *Gene-environment interactions in psychiatry: Joining forces with neuroscience*. Nature Reviews Neuroscience. **7**, 583–590.

Caspi A, Moffitt TE, Cannon M, McClay J, Murray R, Harrington H, Taylor A, Arseneault L, Williams B, Braithwaite A, Poulton R & Craig IW (2005). *Moderation of the effect of adolescent-onset cannabis use on adult psychosis by a functional polymorphism in the C0MT gene: Longitudinal evidence of a gene-environment interaction*. Biological Psychiatry. **57**, 1117–1127.

Caspi A, Sugden K, Moffitt TE, Taylor A, Craig IW, Harrington HL, McClay J, Martin J, Braithwaite A & Poulton R (2003). *Influence of life stress on depression: Moderation by a polymorphism in the 5-HTT gene*. Science. **301**, 386-389.

Chapman S (1996). *Recent advance in tobacco control*. British Medical Journal. **313**, 97-100.

Chlebowski RT, Hendrix SL, Langer RD, Stefanick ML, Gass M, Lane D, Rodabough RJ, Gilligan MA, Cyr MG, Thomson CA, Khandekar J, Petrovitch H, McTiernan A, for the WHI Investigators (2003). *Influence of estrogen plus progestin on breast cancer and mammography in healthy postmenopausal women: the Women's Health Initiative Randomized Trial*. Journal of the American Medical Association. **289**, 3243-3253.

Cochran WG & Chambers SP (1965). *The planning of non-experimental studies of human populations*. Journal of the Royal Statistical Society, Series A (General). **128**, 234-266.

Collaborative Group on Hormonal Factors in Breast Cancer (1997). *Breast cancer and hormone replacement therapy: collaborative reanalysis of data from 51 epidemiological studies of 52,705 women with breast cancer and 108,411 women without breast cancer*. Lancet. **350**, 1047-1059.

Collin J, Lee K & Gilmore A (2003). *Unlocking the corporate documents of British American Tobacco: An invaluable global resource needs radically improved access*. Lancet. **363**, 1746-1747.

Collins R & MacMahon S (2001). *Reliable assessment of the effects of treatment on mortality and major morbidity: I: Clinical trials*. Lancet. **357**, 373-380.

Collins R & MacMahon S (2007). Reliable assessment of the effects of treatments on mortality and major morbidity. In Rothwell PM ed. *Treating individuals: From randomised trials to personalised medicine.* Elsevier. Oxford. 3-35.

Concato JC, Shah N & Horwitz RI (2000). *Randomised, controlled trials, observational studies, and the hierarchy of research designs*. New England Journal of Medicine. **342**, 1887-1892.

Connor J, Rodgers A & Priest P (1999). *Randomised studies of income supplementation: A lost opportunity to assess health outcomes*. Journal of Epidemiology and Community Health. **53**. 725-730.

Cook TD & Campbell DT (1979). *Quasi-experimentation: Design and analysis issues for field settings*. Rand-McNally. Chicago.

Cornfield J (1951). *A method of estimating comparative rates from clinical data. Applications to cancer of the lung, breast, and cervix*. Journal of the National Cancer Institute. **11**, 1269-1275.

Cornfield J, Haenszel W, Hammond E, Lilienfeld A, Shimkin M & Wynder E (1959). *Smoking and lung cancer: Recent evidence and a discussion of some questions*. Journal of the National Cancer Institute. **22**, 173-203.

Costello EJ, Compton FN, Keeler G & Angold A (2003). *Relationships between poverty and psychopathology: A natural experiment*. Journal of the American Medical Association. **290**, 2023-2029.

Dabis F, Msellati P, Dunn D, Lepage P, Newell ML & Perre P (1993). Estimating the rate of mother-to-child transmission of HIV. *Report of a workshop on methodological issues, Ghent, Belgium, 17-20 February, 2002*. AIDS. **7**, 1139-1148.

Davey-Smith G (2006). *Randomized by (your) god: robust inference from a non-experimental study design*. Journal of Epidemiology and Community Health. **60**, 382-388.

Davey-Smith G & Ebrahim S (2003*).'Mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease?* International Journal of Epidemiology. **32**, 1-22.

Dawid AP (2000). *Causal inference without counterfactuals (with Discussion)*. Journal of the American Statistical Association. **95**, 407-448.

Dawid AP (2002). *Influence diagrams for causal modelling and inference*. International Statistical Review. **70**, 161–89. Corrigenda, ibid. 437.

Des Jarlais DC, Lyles C, Crepaz N & the TREND group (2004). *Improving the reporting quality of nonrandomized evaluations of behavioral and public health interventions: The TREND statement*. American Journal of Public Health. **94**, 361-366.

Didelez V & Sheehan N (2005). *Mendelian randomisation and instrumental variables: what can and what can't be done.* http://www.homepages.ucl.ac.uk/~ucakvdi/tech_rep_mendel.pdf

Diez Roux AV, Schwartz S & Susser E (2002). *Ecological variables, ecological studies, and multilevel studies in public health research.* In Detels R, McEwenJ, Beaglehole R & H Tanaka eds. *Oxford Textbook of Public Health, 4th Edition*. Oxford University Press. Oxford. 493-507.

Doll R & Hill AB (1950). *Smoking and carcinoma of the lung: Preliminary report*. British Medical Journal. **2**, 739-748.

Doll R & Hill AB (1954). *The mortality of doctors in relation to their smoking habits. A preliminary report.* British Medical Journal. **2**, 1451-1455.

Doll R, Peto R, Boreham J & Sutherland I (2004). *Mortality in relation to smoking: 50 years' observations on male British doctors*. British Medical Journal. **328**, 1519.

D'Onofrio BM, Turkheimer E, Eaves LJ, Corey LA, Berg K, Solaas MH & Emery RE (2003). *The role of the Children of Twins design in elucidating causal relations between parent characteristics and child outcomes*. Journal of Child Psychology and Psychiatry. **44**, 1130-1144.

D'Onofrio BM, Van Hulle CA, Waldman ID, Rodgers JL, Harden KP, Rathouz PJ & Lahey BB (in press). *Smoking during pregnancy and offspring externalizing problems: An exploration of genetic and environmental confounds*. Development and Psychopathology.

D'Souza Y, Fombonne E & Ward BJ (2006). *No evidence of persisting measles virus in peripheral blood mononuclear cells from children with autism spectrum disorder*. Paediatrics. **118**, 1664-75.

Dunn D, Newell ML, Van Praag E, Ven de Perre P & Peckham C (1992). *Risk of human imunodeficiency virus type 1 transmission through breast feeding*. Lancet. **340**, 585-588.

Eaves LJ (2006). *Genotype x environment interaction in psychopathology: Fact or artifact?* Twin Research and Human Genetics. **9**, 1-8.

Egger M, Davey-Smith, G & Sterne JAC (2002). *Systematic reviews and meta-analysis.* In Detels R, McEwen J, Beaglehole R & Tanaka H eds. *Oxford Textbook of Public Health, 4th Edition*. Oxford University Press. Oxford. 655-675.

Elwood JH & Nevin NC (1973). *Factors associated with anencephalus and spina bifida in Belfast*. British Journal of Preventive and Social Medicine. **27**, 73-80.

Epstein F (1996). *Cardiovascular disease epidemiology*. Circulation. **93**, 1755-1764.

European Collaborative Study (1994*). Caesarean section and risk of vertical transmission of HIV-1 infection*. Lancet. **343**, 1464-1467.

European Mode of Delivery Collaboration (1999). *Elective caesarean section versus vaginal delivery in preventing vertical HIV-1 transmission: A randomized clinical trial*. Lancet. **353**, 1035-1039.

Evans I, Thornton H & Chalmers I (2007). *Testing treatments: Better research for better healthcare.* The British Library. London.

Eysenck HJ (1980). *The causes and effects of smoking.* Maurice Temple Smith. London.

Eysenck HJ (1991). *Smoking, personality and stress: Psychosocial factors in the prevention of cancer and coronary heart disease*. Springer Verlag. New York.

Fergusson DM, Horwood LJ, Caspi A, Moffitt TE & Silva PA (1996). *The (artefactual) remission of reading disability: Psychometric lessons in the study of stability and change in behavioural development*. Developmental Psychology. **32**, 132-140.

Finkelstein MO, Levin B & Robbins H (1966 a). *Clinical and prophylactic trials with assured new treatment for those at greater risk: I: A design proposal*. American Journal of Public Health. **86**, 691-695.

Finkelstein MO, Levin B & Robbins H (1966 b). *Clinical and prophylactic trials with assured new treatment for those at greater risk: II: Examples*. American Journal of Public Health. **86**, 696-705.

Fisher RA (1925). *Statistical methods for research workers*. Oliver & Boyd. Edinburgh.

Fisher RA (1958 a). *Lung cancer and cigarettes*. Nature. **182**, 108.

Fisher RA (1958 b). *Cancer and smoking*. Nature. **182**, 596.

Geoffroy M-C, Côté S, Borge AIH, Larouche F, Séguin JR & Rutter M (2007). *Association between early non-maternal care and children's receptive language skills prior to school entry: the moderating role of the socio-economic status*. Journal of Child Psychology and Psychiatry. **48**, 490-497.

Gigerenzer G (2003). *Reckoning with risk: Learning to live with uncertainty.* Penguin Books. London.

Gilbert R, Salanti G, Harden M & See S (2005). *Infant sleeping position and the sudden infant death syndrome: Systematic review of non-experimental studies and historical review of recommendations from 1940-2002*. International Journal of Epidemiology. **34**, 874-887.

Gill RD & Robins JM (2001). *Causal inference for complex longitudinal data: The continuous case*. Annals of Statistics. **29**, 1785-1811.

Glantz SA, Barnes DE, Bero L, Hanauer P & Slade J (1995). *Looking through a keyhole at the tobacco industry. The Brown and Williamson documents*. Journal of the American Medical Association, **274**, 219-24.

Glasziou P, Chalmers I, Rawlins M & McCulloch P (2007). *When are randomised trials unnecessary? Picking signal from noise*. British Medical Journal. **334**, 349-351.

Grant BF & Dawson DA (1997). *Age at onset of alcohol use and its association with DSMIV alcohol abuse and dependence: Results from the National Longitudinal Alcohol Epidemiologic Survey*. Journal of Substance Abuse. **9**, 103-110.

Gray R & Henderson J (2006). *Review of the fetal effects of prenatal alcohol exposure: Report to the Department of Health.* National Perinatal Epidemiology Unit. Oxford.

Gray RH, Kigozi G, Serwadda D, Makumbi F, Watya S, Nalugoda F, Kiwanuka N, Moulton LH, Chaudhary MA, Chen MZ, Sewankambo NK, Wabwire-Mangen F, Bacon MC, Williams CF, Opendi P, Reynolds SJ Laeyendecker O, Quinn TC & Wawer MJ (2007). *Male circumcision for HIV prevention in men in Rakai, Uganda: A randomised trial*. Lancet. **369**, 657-666.

Green ML, Singh AV, Zhang Y, Nemeth KA, Sulik KK & Knudsen TB (2007). *Reprogramming of genetic networks during initiation of the Fetal Alcohol Syndrome*. Developmental Dynamics. **236**, 613-631.

Greenland S (1992). *Divergent biases in ecologic and individual-level studies*. Statistics in Medicine. **11**, 1209-1223.

Greenland S (2005). *Multiple-bias modelling for analysis of observational data (with discussion)*. Journal of the Royal Statistical Society: Series A (Statistics in Society). **168**, 267–306.

Gregg NMcA (1941). *Congenital cataract following German measles in mother*. Transactions of the Ophthalmological Society of Australia. **3**, 35-46.

Grodstein F, Manson JE & Stampfer MJ. (2001). *Postmenopausal hormone use and secondary prevention of coronary events in the Nurses' Health Study*. Annals of Internal Medicine. **135**, 1-8.

Grodstein F, Manson JE & Stampfer MJ (2006). *Hormone therapy and coronary heart disease: The role of time since menopause and age at hormone initiation*. Journal of Women's Health. **15**, 35-44.

Gunnar MR & Vasquez M (2006). *Stress neurobiology and developmental psychopathology*. In Cicchetti D & Cohen DJ eds. *Developmental Psychopathology, 2nd Edition. Vol. 2: Developmental Neuroscience*. Wiley. New York. 533-577.

Halperin DT & Bailey RC (1999). *Male circumcision and HIV infection: 10 years and counting.* Lancet. **354**, 1813-1815.

Hanshaw JB, Dudgeon JA & Marshall WC (1985). *Viral diseases of the fetus and newborn. 2nd Edn*. W.B. Saunders. Philadelphia.

Harbst AL, Ulfelder H & Poskanzer DC (1971). *Adenocarcinoma of the vagina. Association of maternal stilbestrol therapy with tumour appearance in young women*. New England Journal of Medicine. **284**, 878-881.

Hariri AR, Mattay VS, Tessitore A, Kolachana B, Fera F, Goldman D, Egan MF & Weinberger DF (2002). *Serotonin transporter genetic variation and the response of the human amygdala*. Science*.* **297***,* 400-404.

Heckman JJ & Smith JA (1995). *Assessing the case for social experiments*. Journal of Economic Perspectives. **9**, 85-110.

Heckman J & Vytlacil E (1999). *Local instrumental variables and latent variable models for identifying and bounding treatment effects*. Proceedings of the National Academy of Science. **96**, 4730–4734.

Heckman JJ (In Press). *Randomised evaluations.* In Heckman JJ ed. *Handbook of Econometrics*.

Hemminki E & McPherson K (2000). *Value of drug-licensing documents in studying the effect of postmenopausal hormone therapy on cardiovascular disease*. Lancet. **355**, 566-569.

Hernán MA, Herndandez-Diaz S & Robins JM (2004). *A structural approach to selection bias*. Epidemiology. **15**, 615-625.

Hernán M & Robins J (2006). *Instruments for causal inference. An epidemiologist's dream?* Epidemiology. **17**, 360–372.

Hernán MA, Robins JM & Rodriguez LAG (2005). *Discussion of Prentice et al.* Biometrics. **61**, 922-929.

Hernandez LM & Blazer DG eds. (2006). *Genes, behaviour, and the social environment: Moving beyond the nature-nurture debate.* National Academies Press. Washington DC.

Hibbard ED & Smithells RW (1965). *Folic acid metabolism and human embryopathy*. Lancet. **1**, 1254.

Hill AB (1965). *The environment and disease: Association or causation?* Proceedings of the Royal Society of Medicine. **58**, 295-300.

Hilts J (1996). *Smokescreen: The truth behind the tobacco industry cover-up.* Addison-Wesley. Reading, MA.

Honda H, Shimizu Y & Rutter M (2005). *No effect of MMR withdrawal on the incidence of autism: a total population study.* Journal of Child Psychology and Psychiatry. **46**, 572-579.

House of Commons Science and Technology Committee (2006). *Scientific advice, risk and evidence-based policy making.* The Stationery Office Ltd. London.

Hotz VJ, Imbens GW & Klerman JA (2006). *Evaluating the differential effects of alternative welfare-to-work training components: A reanalysis of the California GAIN program*. Journal of Labor Economics. **24**, 521-567.

Hsia J, Criqui MH, Herrington DM, Manson JE, Wu L, Heckbert SR, Allison M, McDermott MM, Robinson J, Masaki K & Women's Health Initiative Research Group. (2006). *Conjugated equine estrogens and coronary heart disease: The Women's Health Initiative*. Archives of Internal Medicine. **166**, 357–365.

Husan B & Bhutta ZA (2007). *Periconceptual supplementation for preventing neural tube defects*. WHO reproductive health library http://www.rhlibrary.com/Commentaries/htm/Bhcom.htm

Imbens G & Angrist J (1994). *Identification and estimation of local average treatment effects.* Econometrica. **62**, 467–75.

Irons DE, McGue M, Iacono WG & Oetting WS (2007). *Mendelian Randomization: A novel test of the gateway hypothesis and models of gene-environment interplay*. Development and Psychopathology. **19,** 1181-1195.

Jaffee SR, Caspi A, Moffitt TE, Polo-Tomas M, Price TS, & Taylor A (2004). *The limits of child effects: Evidence for genetically mediated child effects on corporal punishment but not on physical maltreatment*. Developmental Psychology. **40**, 1047-1058.

Jaffee SR, Moffitt TE, Caspi A & Taylor A (2002). *Life with (and without) father: The benefits of living with two biological parents depend on the father's antisocial behaviour*. Child Development. **41**, 1095-1103.

Joffe MM & Rosenbaum PR (1999). *Invited commentary: Propensity scores*. American Journal of Epidemiology. **150**, 327–33.

Jones KL, Smith DW, Ulleland CH & Streissguth AP (1973). *Pattern of malformation in offspring of chronic alcoholic mothers*. Lancet. **1**, 1267-1271.

Jones PB & Fung WLA (2005). *Ethnicity and mental health: The example of schizophrenia in the African-Caribbean population in Europe*. In Rutter M & Tienda M eds. *Ethnicity and causal mechanisms*. Cambridge University Press. New York. 227-261.

Katan MB (1986). *Apolipoprotein E isoforms, serum cholesterol, and cancer*. Lancet. **327**, 507-508.

Keavney B, Danesh J, Paris S, Palmer A, Clark S, Youngman L, Delépine M, Lathrop M, Peto R & Collins R (2006). *Fibrinogen and coronary heart disease: Test of causality by 'Mendelian randomization'*. International Journal of Epidemiology. **35**, 935-943.

Kendler KS & Prescott CA (2006). *Genes, environment, and psychopathology: Understanding the causes of psychiatric and substance use disorders.* Guilford Press. New York.

Keys A (1980). *Seven countries: A multivariate analysis of death and coronary heart disease*. Harvard University Press. Cambridge, MA.

Khan MA, Herzog CA, St Peter JV, Hartley GG, Madlon-Kay R, Dick CD, Asinger RW & Vessey JT (1998) *The prevalence of cardiac valvular insufficiency assessed by transthoracic echocardiography in obese patients treated with appetite-suppressant drugs*. New England Journal of Medicine. **339**, 713-718.

Kim-Cohen J, Caspi A, Taylor A, Williams B, Newcombe R, Craig IW & Moffitt TE (2006). *MAOA, maltreatment, and gene-environment interaction predicting children's mental health: New evidence and a meta-analysis*. Molecular Psychiatry. **11**, 903-913.

Kim-Cohen J, Moffitt TE, Taylor A, Pawlby SJ & Caspi A (2005). *Maternal depression and children's antisocial behavior: Nature and nurture effects*. Archives of General Psychiatry. **62**, 173-181.

Kunz R & Oxman AD (1998). *The unpredictability paradox: Review of empirical comparisons of randomised and non-randomised clinical trials*. British Medical Journal. **317**, 1185-1190.

Kurth T, Walker A, Glynn R, Chan K, Gaziano J, Berger K & Robins J (2006). *Results of multivariable logistic regression, propensity matching, propensity adjustment, and propensity-based weighting under conditions of nonuniform effect*. American Journal of Epidemiology. **163**, 262–70.

Laird NM & Mosteller F (1990). *Some statistical methods for combining experimental results*. International Journal of Technology Assessment in Health Care. **6**, 5-30.

Landovitz RJ (2007). *Recent efforts in biomedical prevention of HIV*. Topics in HIV Medicine. **15**, 99-103.

Lawlor DA, Davey-Smith G & Ebrahim S (2004). *Socioeconomic position and hormone replacement therapy use: Explaining the discrepancy in evidence from observational and randomized controlled trials*. American Journal of Public Health. **94**, 2149-2154.

Le Fanu J *(1999). The rise and fall of modern medicine.* Abacus. London.

Lenz W *(1962). Thalidomide and congenital abnormalities. Lancet.* **1***,* 271-272.

Leventhal T & Brooks-Gunn J (2004). *A randomized study of neighborhood effects on low-income children's educational outcomes*. Developmental Psychology. **40**, 488-507.

Leventhal T, Fauth RC & Brooks-Gunn J (2005). *Neighborhood poverty and public policy: A five-year follow-up of children's educational outcomes in the New York City Moving to Opportunity demonstration*. Developmental Psychology. **41**, 933-952.

Llewelyn CA, Hewitt PE, Knight RSG, Amar K, Cousens S, Mackenzie J & Will RG (2004). *Possible transmission of variant Creutzfeld-Jakob disease by blood transfusion.* Lancet. **363**, 417-421.

Ludwig J & Miller DL (2007). *Does Head Start improve children's life chances? Evidence from a regression discontinuity design*. Quarterly Journal of Economics. **122**, 159-208.

Luellen JK, Shadish WR & Clark MH (2005). *Propensity scores: An introduction and experimental test*. Evaluation Review. **29**, 530-558.

Lykken DT (1968). *Statistical significance in psychological research*. Psychological Bulletin. **70**, 151-159.

Lynch SK, Turkheimer E, D'Onofrio BM, Mendle J & Emery RE (2006). *A genetically informed study of the association between harsh punishment and offspring behavioural problems*. Journal of Family Psychology. **20**, 190-198.

Mackenbach JP (2002). *Socio-economic inequalities in health in developed countries: the facts and the options.* In Detels R, McEwen J, Beaglehole R & Tanaka H eds. *Oxford Textbook of Public Health, 4th Edition*. Oxford University Press. Oxford. 1773-1790.

MacKenzie R, Palmer CR, Lomas DJ, & Dixon AK (1996). *Magnetic resonance imaging of the knee: Assessment of effectiveness*. Clinical Radiology. **51**, 245-250.

Mackie JL (1965). *Causes and conditions*. American Philosophical Quarterly. **2**, 245-264.

Mackie JL (1974). *The cement of the universe: A study of causation.* Oxford University Press. Oxford.

MacMahon S & Collins R (2001). *Reliable assessment of the effects of treatment on mortality and major morbidity: II Non-experimental studies*. Lancet. **357**, 455-462.

MacMahon B, Pugh T & Ipsen J (1960). *Epidemiologic methods.* Little Brown. Boston.

Madsen K, Lauritsen MB, Pedersen CB, Thorsen P, Plesner A-M, Andersen PH & Mortensen PB (2003). *Thimerosal and the occurrence of autism: Negative ecological evidence from Danish population-based data.* Pediatrics. **112**, 604-606.

Maldonado G & Greenland S (2002). *Estimating causal effects. International Journal of Epidemiology.* **31**, 422-429.

Mamdani M, Sykora K, Li P, Normand S-LT, Streiner DL, Austin PC, Rochon PA & Anderson GM (2005). *Reader's guide to critical appraisal of cohort studies: 2. Assessing potential for confounding*. British Medical Journal. **330**, 960-962.

Manson JE, Hsia J, Johnson KC, Rossouw JE, Assaf AR, Lasser NL, Maurizio Trevisan M, Black HR, Heckbert SR, Detrano R, Strickland OL, Wong ND, Crouse JR, Stein E, Cushman M for the Women's Health Initiative Investigators (2003). *Estrogen plus progestin and the risk of coronary heart disease*. The New England Journal of Medicine. **349**, 523-534.

Manson JE & Martin KA (2001). *Clinical practice: Postmenopausal hormone replacement therapy*. New England Journal of Medicine. **345**, 34-40.

March D & Susser E (2006). *The eco- in eco-epidemiology.* International Journal of Epidemiology. **35**, 1379-1383.

Marmot MG (2004). *Status Syndrome: How your social standing directly affects your health and life expectancy.* Bloomsbury Publishing Plc. London.

Marmot MG, Bosma H, Hemingway H, Brunner E & Stansfeld S (1997). *Contribution of job control and other risk factors to social variations in coronary heart disease incidence*. Lancet. **350**, 235-239

Marmot MG & Syme SL (1976). *Acculturation and coronary heart disease in Japanese-Americans*. American Journal of Epidemiology. **104**, 225-247.

Marmot MG & Wilkinson R eds. (2006). *Social determinants of health, 2nd Edition*. Oxford University Press. Oxford.

Martin TR & Bracken MB (1987). *The association between low birth weight and caffeine consumption during pregnancy*. American Journal of Epidemiology. **126**, 813-821.

McBride WG (1961). *Thalidomide and congenital abnormalities*. Lancet. **2**, 1358.

McClellan JM, Susser E & King M-C (2006). *Maternal famine, de novo mutations, and schizophrenia*. Journal of the American Medical Association. **296**, 582-584.

McPherson K (2004). *Where are we now with hormone replacement therapy*? British Medical Journal. **328**, 357-358.

McPherson K, Britton AR & Wennberg JE (1997). *Are randomized controlled trials controlled? Patient preferences and unblind trials*. Journal of the Royal Society of Medicine. **90**, 652-656.

Meade TW, Humphries SE & De Stavola BL (2006). *Commentary: Fibrinogen and coronary heart disease – test of causality by 'Mendelian' randomization by Keavney et al.* International Journal of Epidemiology. **35**, 944-947.

Medical Research Council Streptomycin in Tuberculosis Trials Committee (1948). *Streptomycin treatment for pulmonary tuberculosis*. British Medical Journal. **2**, 769-782.

Medical Research Council (1949). *Treatment of pulmonary tuberculosis with PAS and streptomycin: Preliminary report*. British Medical Journal. 31.12.49, 1521.

Medical Research Council Vitamin Study Research Group (1991). *Prevention of neural tube defects: Results of the Medical Research Council Vitamin Study*. Lancet. **338**, 131-137.

Meyer-Lindenberg A, Buckholtz JW, Kolachana B, Hariri AR, Pezawas L, Blasi G, Wabnitz A, Honea R, Verchinski BA, Callicott J, Egan MF, Mattay VS & Weinberger DR (2006). *Neural mechanisms of genetic risk for impulsivity and violence in humans*. Proceedings of the National Academy of Sciences of the USA. **103**, 6269-6274.

Michels KB & Manson JE (2003). *Postmenopausal hormone therapy: A reversal of fortune*. Circulation. **107**, 1830-1833.

Miech RA, Caspi A, Moffitt TE, Entner BR & Silva PA (1999). *Low socio-economic status and mental disorders: A longitudinal study of selection and causation during young adulthood*. American Journal of Sociology. **104**, 1096-1131.

Mill JS (1843). *A system of logic.* Parker. London.

Mill JL, Rhoads GG, Simpson JL, Cunningham GC, Conley MR, Lassman MR, Walden ME, Depp OR & Hoffman HJ (1989). *The absence of a relation between the periconceptional use of vitamins and neural-tube defects. National Institute of Child Health and Human Development Neural Tube Defects Study Group*. New England Journal of Medicine. **321**, 430-435.

Millen JW (1962). *Thalidomide and limb deformities*. Lancet. **2**, 599-600.

Million Women Study Collaborators (2003). *Breast cancer and hormone replacement therapy in the Million Women Study*. Lancet. **362**, 419-427.

Million Women Study Collaborators (2007). *Ovarian cancer and hormone replacement therapy in the Million Women Study*. Lancet. **369**, 1703-1710.

Milunsky A, Jick H, Jick SS, Bruell CL, MacLoughlin DS, Rothman KJ & Willett W (1989). *Multivitamin/folic acid supplementation in early pregnancy reduces the prevalence of neural tube defects*. Journal of the American Medical Association. **262**, 2847-2852.

Minelli C, Abrams KK, Sutton AJ & Cooper NJ (2004). *Benefits and harms associated with hormone replacement therapy: Clinical decision analysis*. British Medical Journal. **328**, 371-375.

Moe V (2002). *Foster-placed and adopted children exposed in utero to opiates and other substances: prediction and outcome at four and a half years*. Journal of Developmental and Behavioural Paediatrics. **23**, 330-339.

Moffitt TE, Caspi A & Rutter M (2005). *Strategy for investigating interactions between measured genes and measured environments*. Archives of General Psychiatry. **62**, 473-481.

Moore THM, Zammit S, Lingford-Hughes A, Barnes TRE, Jones PB, Burke M & Lewis G (2007). *Cannabis use and risk of psychotic or affective mental health outcomes: a systematic review*. Lancet. **370**, 319-328.

Moses S, Bailey RC & Ronald AR (1998). *Male circumcision: Assessment of health benefits and risks*. Sexually Transmitted Infections. **74**, 368-373.

Murch SH, Anthony A, Casson DH, Malik M, Berelowitz M, Dhillon AP, Thomson MA, Valentine A, Davies SE & Walker-Smith JA (2004). *Retraction of an interpretation*. Lancet. **363**, 750.

Murray CJ, Lauer JA, Hutubessy RC, Niessen L, Tomijima N, Rodgers A, Lawes CM & Evans DB (2003). *Effectiveness and costs of interventions to lower systolic blood pressure and cholesterol: a global and regional analysis on reduction of cardiovascular-disease risk*. Lancet. **361**, 717-725.

Nduati R, John G, Ngacha DA, Richardson S, Overbaugh J, Mwatha A, Ndinya-Achola J, Bwayo J, Onyango FE & Kreiss J (2000). *Effect of breastfeeding and formula feeding on transmission of HIV-1: A randomized clinical trial*. Journal of the American Medical Association. **283**, 1167-1174.

Nelson CA & Jeste S (in press). *Neurobiological perspectives on developmental psychopathology.* In: Rutter M, Bishop D, Pine D, Scott S, Stevenson J, Taylor E & Thapar A eds. *Rutter's Child and Adolescent Psychiatry, 5th Edition.* Blackwell. Oxford.

NICE (2007). *The guidelines manual 2007*. http://www.nice.org.uk/page.aspx?o=422950

Nitsch D, Molokhia M, Smeeth L, DeStavola BL, Whittaker JC & Leon DA (2006). *Limits to causal inference based on Mendelian randomization: A comparison with randomized controlled trials*. American Journal of Epidemiology. **163**, 397-403.

Normand ST, Sykora K, Li P, Mamdani M, Rochon PA & Anderson GM (2005). *Readers' guide to critical appraisal of cohort studies: 3. Analytical strategies to reduce confounding*. British Medical Journal. **330**, 1021–1023.

Nuffield Council on Bioethics (2005). *The ethics of research involving animals*. Nuffield Council on Bioethics. London.

O'Connor TG, Deater-Deckard K, Fulker D, Rutter M & Plomin R (1998). *Genotype-environment correlations in late childhood and early adolescence: Antisocial behavioral problems and coercive parenting*. Developmental Psychology. **34**, 970-981.

Office of the US Surgeon General (1964). *Smoking and health: Report of the Advisory Committee of the US Surgeon General.* Office of the Surgeon General. Washington DC.
Office of the US Surgeon General (2004). *The Health Consequences of Smoking*. US Department of Health and Human Services. Atlanta.

Office of the US Surgeon General (2006). *Surgeon General's Report – The Health Consequences of Involuntary Exposure to Tobacco Smoke*. US Department of Health and Human Services. Atlanta.

Ong EK & Glantz SA (2000). *Tobacco industry efforts subverting International Agency for Research on Cancer's second-hand smoke study*. Lancet. **355**, 1253-1259.

Padwal R, Straus SE & McAlister FA (2007). *Cardiovascular risk factors and their effects on the decision to treat hypertension: evidence based review*. British Medical Journal. **322**, 977-980.

Pahor M, Psaty BM, Alderman MH, Applegate WB, Williamson JD, Cavazzini C & Furberg CD (2000). *Health outcomes associated with calcium antagonists compared with other first-line antihypertensive therapies: a meta-analysis of randomised controlled trials*. Lancet. **356***,* 1949-1954.

Paling J (2003). *Strategies to help patients understand risks*. British Medical Journal. **327**, 745-748.

Parkman PD, Buescher EL & Artenstein MS (1962). *Recovery of rubella virus from army recruits*. Proceedings of the Society for Experimental Biology and Medicine. **111**, 225-230.

Pearl J (1995). *Causal diagrams for empirical research*. Biometrika. **82**, 669-688.

Pearl J (2000). *Causality: Models, reasoning, and inference*. Cambridge University Press. Cambridge, UK.

Peden AH, Head MW, Ritchie DL, Bell JE & Ironside JW (2004). *Preclinical cGJD after blood transfusion in a PRNP codon 129 heterozygous patient*. Lancet. **364**, 527-529.

Petitti DB & Freedman DA (2005). *Invited commentary: How far can epidemiologists get with statistical adjustment?* American Journal of Epidemiology. **162**, 415-418.

Peto R, Darby S, Deo H, Silcocks P, Whitley E & Doll R (2000). *Smoking, smoking cessation, and lung cancer in the UK since 1950: combination of national statistics with two case-control studies*. British Medical Journal. **321**, 323-329.

Phillips CV & Goodman KJ (2004). *The missed lessons of Sir Austin Bradford Hill*. Epidemiologic Perspectives & Innovations. 1, doi: 10.1186/1742-5573-1-2.

Pickles A (1998). Generalized estimating equations. In Armitage P & Colton T eds. *The Encyclopedia of Biostatistics*. Wiley. New York. 1626–1637.

Pickles A (in press). *What clinicians need to know about statistical issues and methods.* In Rutter M, Bishop D, Pine D, Scott S, Stevenson J, Taylor E & Thapar A eds. *Rutter's Child and Adolescent Psychiatry, 5th Edition.* Blackwell. Oxford.

Platt JR (1964). *Strong inference*. Science. **146**, 347-353.
Plomin R & Bergeman CS (1991). *The nature of nurture: Genetic influence on 'environmental' measures*. Behavioral and Brain Sciences. **14**, 373-386.

Plomin R, Owen MJ & McGuffin P (1994). *The genetic basis of complex human behaviors*. Science. **264**, 1733-1739.

Pocock SJ & Elbourne DR (2000). *Randomized trials or observational tribulations?* New England Journal of Medicine. **342**, 1907-1909.

Popper K (1959). *The logic of scientific discovery*. Hutchinson. London.

Prentice RL, Langer R, Stefanick ML, Howard BV, Pettinger M, Anderson G, Barad D, Curb JD, Kotchen J, Kuller L, Limacher M & Wactawski-Wende J for the Women's Health Initiative Investigators. (2005 a). *Combined postmenopausal hormone therapy and cardiovascular disease: Toward resolving the discrepancy between non-experimental studies and the Women's Health Initiative clinical trial*. American Journal of Epidemiology. **162**, 404-414.

Prentice RL, Pettinger M & Anderson GL (2005 b). *Statistical issues arising in the Women's Health Initiative*. Biometrics. **61**, 899-941.

Prentice RL, Langer R, Stefanick ML, Howard, BV Pettinger M, Anderson G, Barad D, Curb JD, Kotchen J, Kuller L, Limacher M & Wactawski-Wende J for the Women's Health Initiative Investigators. (2006). *Combined analysis of Women's Health Initiative non-experimental and clinical trial data on postmenopausal hormone treatment and cardiovascular disease*. American Journal of Epidemiology. **163**, 589-599.

Prescott CA & Kendler KS (1999). *Age at first drink and risk for alcoholism: A noncausal association*. Alcoholism: Clinical and Experimental Research. **23**, 101-107.

Psaty BM & Furberg CD (2004). *Contemplating ACTION – long-acting nifedipine in stable angina*. Lancet. **364**, 817-818.

Psaty BM, Heckbert SR, Koepsell TD, Siscovick DS, Raghunathan TE, Weiss NS, Rosendaal FR, Lemaitre RN, Smith NL, Wahl PW, Siscovick DS & Wagner EH (1995). *The risk of myocardial infarction associated with antihypertensive drug therapies*. Journal of the American Medical Association. **274**, 620-625.

Pulkkinen L, Kaprio J & Rose R (2006). *Socioemotional development and health from adolescence to adulthood.* Cambridge University Press. Cambridge.

Rahaman MM, Aziz KM, Patwari Y & Munshi MH (1979). *Diarrhoeal mortality in two Bangladeshi villages with and without community-based oral rehydration therapy*. Lancet. **2(8147)**, 809-812.

Raiffa H (1968). *Decision Analysis*. Addison-Wesley. Reading, MA.

Randall CL (2001). *Alcohol and pregnancy: highlights from three decades of research*. Journal of Studies on Alcohol. **62**, 554-561.

Reichart CS & Gollob HF (1986). *Satisfying the constraints of causal modelling*. In Trochim WMK ed. *Advances in quasi-experimental design and analysis*. Jossey-Bass. San Francisco. 91-107.

Reiss Jr AJ (1995). *Community influences on adolescent behavior.* In Rutter M ed. *Psychosocial disturbances in young people: Challenges for prevention.* Cambridge University Press. Cambridge. 305-332.

Robins JM (2001). *Data, design and background knowledge in etiologic inference*. Epidemiology. **11**, 313-320.

Robins JM & Greenland S (1989). *Identification of causal effects using instrumental variables: Comment*. Journal of the American Statistical Association. **91**, 456-458.

Robins JM, Hernán MA & Brumback B (2000). *Marginal structural models and causal inference in epidemiology.* Epidemiology. **11**, 550-560.

Robinson WS (1950). *Ecological correlations and the behaviour of individuals*. American Sociology Review. **15**, 351-357.

Rochon PA, Gurwitz JH, Sykora K, Mamdani M, Steiner DL, Garfinkel S, Normand S-LT & Anderson GM (2005). *Reader's guide to critical appraisal of cohort studies: 1. Role and design*. British Medical Journal. **330**, 895-897.

Rosenbaum PR (2001). *Stability in the absence of treatment*. Journal of the American Statistical Science Association. **96**, 210-219.

Rosenbaum PR (2002). *Observational Studies*, 2nd edn. Springer-Verlag. New York.

Rosenbaum PR (2007). *Five suggestions for non-experimental studies of treatment effects.* Evidence to working party.

Rosenbaum PR & Rubin DB (1983). *The central role of the propensity score in observational studies for causal effect*. Biometrika. **70**, 41–55.

Rosenbaum PR & Rubin DB (1993 a). *Assessing sensitivity to an unobserved binary covariate in an non-experimental study with binary outcome*. Journal of the Royal Statistical Society, Series B (Methodological). **45**, 212-218.

Rosenbaum PR & Rubin DB (1993 b). *The central role of the propensity score in non-experimental studies for causal effects*. Biometrika. **70**, 41-55.

Rossouw JE, Anderson GL, Prentice RL, LaCroix AZ, Kooperberg C, Stefanick ML, Jackson RD, Beresford SA, Howard BV, Johnson KC, Kotchen JM, Ockene J & Writing Group for the Women's Health Initiative Investigators (2002). *Risks and benefits of estrogen plus progestin in healthy postmenopausal women: Principal results from the Women's Health Initiative randomized controlled trial*. Journal of the American Medical Association. **288**, 321-333.

Rossouw JE, Prentice RL, Manson DE, Wu L-L, Barad D, Barnabei V, Ko M, LaCroix AZ, Margolis KL & Stefanick ML (2007). *Postmenopausal hormone therapy and risk of cardiovascular disease by age and years since menopause*. Journal of the American Medical Associatio. **297**, 1465-1477.

Rothman KJ & Greenland S (1998). *Modern epidemiology, 2nd Edition*. Lippincott-Raven. Philadelphia.

Rothman KJ & Greenland S (2002). Causation and causal inference. In: Detels R, McEwen J, Beaglehole R & Tanaka H eds. *Oxford textbook of Public Health, 4th Edn.* Oxford University Press. Oxford. 641-653.

Royal College of Physicians (1962). *Smoking and health*. Pitman Medical Publishing Co. Ltd. London.

Royal College of Physicians and ASH (2004). *Forty Fatal Years*. RCP and ASH. London.

Royal Institution of Great Britain, SIRC & the Royal Society (2001). *Guidelines on science and health communication*. The Royal Institution, SIRC & The Royal Society. London.

Royal Society (2006). *Science in the public interest*. http://www.royalsoc.ac.uk/downloaddoc.asp?id=2879

Rubin DB (1974). *Estimating causal effects of treatments in randomized and nonrandomized studies*. Journal of Educational Psychology. **66**, 688–701.

Rubin DB (1977). *Assignment to treatment group on the basis of a covariate*. Journal of Educational Statistics. **2**, 4-58.

Rubin DB (1978). *Bayesian inference for causal effects: the role of randomization*. Annals of Statistics. **6**, 34–68.

Rubin DB (1979). *Using multivariate matched sampling and regression adjustment to control bias in observational studies*. Journal of the American Statistical Association. **74**, 318-328.

Rubin DB (1986*). Statistics and causal inference: Comment: Which ifs have causal answers*. Journal of the American Statistical Association. **81**, 961-962.

Rubin DB (2004). *Direct and indirect causal effects via potential outcomes (with discussion)*. Scandinavian Journal of Statistics. **31**, 161–170, 189–198.

Rubin DB (2007). *The design versus the analysis of observational studies for causal effects: Parallels with the design of randomized trials*. Statistics in Medicine. **26**, 20–36.

Rutter M (1971). *Parent-child separation: Psychological effects on the children*. Journal of Child Psychology & Psychiatry. **12**, 233-260.

Rutter M (1974). Epidemiological strategies and psychiatric concepts in research on the vulnerable child. In Anthony EJ & Koupernik C eds. *The child in his family: Children at psychiatric risk.* John Wiley & Sons. New York. 167-179.

Rutter M (1983). *The relationship between science and policy making: The case of lead*. Clean Air. **13**, 17-32.

Rutter M (1998). *Routes from research to clinical practice in child psychiatry: Retrospect and prospect*. Journal of Child Psychology and Psychiatry. **39**, 805-816.

Rutter M (2005). *Incidence of autism spectrum disorders: changes over time and their meaning*. Acta Paediatrica. **94**, 2–15.

Rutter M (2006). Genes and behaviour: Nature-nurture interplay explained. Blackwell Publishing. London.

Rutter M (2007 a). *Gene-environment interdependence*. Developmental Science. **10**, 12-18.

Rutter M (2007 b). *Proceeding from observed correlation to causal inference: The use of natural experiments*. Perspectives on Psychological Science. **2**, *377-395.*

Rutter M (2007 c) SureStart Local Programmes: an ousider's perspective. In Belsky J, Barnes J & Melhuish E eds. *The National Evaluation of SureStart: Does area-based early intervention work?* Policy Press. Bristol. 197-210.

Rutter M, Beckett C, Castle J, Colvert E, Kreppner J, Mehta M, Stevens SE & Sonuga-Barke EJS (2007). *Effects of profound early institutional deprivation: an overview of findings from a UK longitudinal study of Romanian adoptees*. European Journal of Developmental Psychology. **4**, 332-350.

Rutter M, Moffitt TE & Caspi A (2006). *Gene-environment interplay and psychopathology: multiple varieties but real effects*. Journal of Child Psychology and Psychiatry. **47**, 226-261.

Rutter M, Pickles A, Murray R & Eaves L (2001). *Testing hypotheses on specific environmental causal effects on behavior*. Psychological Bulletin. **127**, 291-324.

Rutter M, & Russell Jones R eds. (1983). *Lead versus health: sources and effects of low level lead exposure*. Wiley. Chichester.

Rutter M & Silberg J (2002). *Gene-environment interplay in relation to emotional and behavioural disturbance*. Annual Review of Psychology. **53**, 463-490.

Rutter M, Thorpe K, Greenwood R, Northstone K & Golding J (2003). *Twins as a natural experiment to study the causes of mild language delay: I. Design; twin-singleton differences in language, and obstetric risks*. Journal of Child Psychology and Psychiatry. **44**, 326-334.

Rzany B, Correia O, Kelly JP, Naldi L, Auquier A & Stern R for the Study Group of the International Case-Control Study on Severe Cutaneous Adverse Reactions (1999). *Risk of Stevens-Johnson syndrome and toxic epidermal necrolysis during first weeks of antiepileptic therapy: A case-control study*. Lancet. **353**, 2190-2194.

Sampson RJ & Laub JH (1993). *Crime in the making: Pathways and turning points through life*. Harvard University Press. Cambridge, MA.

Sampson RJ & Laub JH (1996). *Socioeconomic achievement in the life course of disadvantaged men: Military service as a turning point, circa 1940-1965*. American Sociology Review. **61**, 347-367.

Sampson RJ, Laub JH & Wimer C (2006). *Does marriage reduce crime? A counterfactual approach to within-individual causal effects*. Criminology. **44**, 465-508.

Sampson RJ, Raudenbush SW & Earls F (1997). *Neighborhoods and violent crime: A multilevel study of collective efficacy*. Science. **277**, 918-924.

Scargle JD (2000). *Publication bias: The 'File-Drawer' problem in scientific inference*. Journal of Scientific Exploration. **14**, 94-106.

Scherer RW & Langenberg P & von Elm E (2007). *Full publication of results initially presented in abstracts*. Cochrane Database of Systematic Reviews. Issue 2. Art. No.: MR000005. DOI: 10.1002/14651858.MR000005.pub3.

Schwartz J (1994). *Low-level lead exposure and children's IQ: A meta-anaylsis and search for a threshold*. Environmental Research. **65**, 42-55.

Schwartz S & Susser E (2006). *What is a cause?* In Susser E, Schwartz S, Morabia A & Bromet EJ. *Psychiatric epidemiology: Searching for the causes of mental disorders.* Oxford University Press. Oxford & New York. 33-42.

Shadish WR & Cook TD (1999). *Design rules: More steps towards a complete theory of quasi-experimentation*. Statistical Science. **14**, 294-300.

Shadish WR, Cook TD & Campbell DT (2002). *Experimental and quasi-experimental designs for generalized causal inference.* Houghton Mifflin Company. Boston & New York.

Shadish WR, Luellen JK & Clark MH (2006). Propensity scores and quasi-experiments: A testimony to the practical side of Lee Sechrest. In Bootzin RR & Mcknight PE eds. *Strengthening research methodology: Psychological measurement and evaluation.* American Psychological Association. Washington DC. 143-157.

Shavelson RJ & Towne L eds. (2002). *Scientific research in education.* National Academy Press. Washington DC.

Silberg JL & Eaves LJ (2004). *Analysing the contributions of genes and parent-child interaction to childhood behavioural and emotional problems: a model for the children of twins.* Psychological Medicine. **34**, 347-356.

Simpson SH, Eurich DT, Majumdar SR, Padwal RS, Tsuyuki RT, Varney J & Johnson JA (2006). *A meta-analysis of the association between adherence to drug therapy and mortality*. British Medical Journal. **333**, 15-20.

Singer LT, Minnes S, Short E, Arendt R, Farkas K, Lewis B, Klein N, Russ S, Min MO & Kirchner HL (2004). *Cognitive outcomes of preschool children with prenatal cocaine exposure*. Journal of the American Medical Association. **291**, 2448-2456.

SIRC and ASCOR (2006). *Messenger Report: Media, science and society. Engagement and Governance in Europe*. SIRC. Oxford.

Smithells RW, Shephard S, Schorah CJ, Seller MJ, Nevin NC, Harris R, Read AP & Fielding DW (1980). *Possible prevention of neural tube defects by periconceptual vitamin supplementation*. Lancet. **1**, 339-340.

Solberg TK, Nygaard ØP, Sjaavik K, Hofoss D & Ingebrigtsen T (2005). *The risk of 'getting worse' after lumbar microdiscectomy*. European Spine Journal. **14**, 49-54.

Sonuga-Barke EJS, Beckett C, Krepppner J, Castle J, Colvert E, Stevens S, Hawkins A & Rutter M (submitted). *Is subnutrition necessary for a poor outcome following severe and pervasive early institutional deprivation? Brain growth, cognition and mental health.*

Spiegelhalter D, Freedman L & Parmar M (1994). *Bayesian approaches to randomized trials*. Journal of the Royal Statistical Society Series. **A157**, 357–416.

Spira A, Beane J, Shah V, Liu G, Schembri F, Yang X., Palma J & Brody JS (2004). *Effects of cigarette smoke on the human airway epithelial cell transcriptome*. Proceedings of the National Academy of Sciences of the United States of America. **101**, 10143-10148.

Spitzer WO, Aitken KJ, Dell'Aniello S & Davis MW (2001). *The natural history of autistic syndrome in British children exposed to MMR*. Adverse Drug Reactions and Toxicological Reviews. **20**, 160-163.

Stampfer MJ & Colditz GA (1991). *Estrogen replacement therapy and coronary heart disease: A quantitative assessment of the epidemiologic evidence*. Preventive Medicine. **20**, 47-63.

St. Clair D, Xu M, Wang P, Yu Y, Fang Y, Zhang F, Sheng X, Gu N, Feng G, Sham P & He L (2005). *Rates of adult schizophrenia following prenatal exposure to the Chinese famine of 1959-1961*. Journal of the American Medical Association. **294**, 557-562.

Stattin H & Magnusson D (1990). *Paths through life: Vol. 2. Pubertal maturation in female development.* Erlbaum. Hillsdale, NJ.

Stein ZA, Susser M, Saenger G & Marolla F (1975). *Famine and human development: The Dutch hunger winter of 1944-1945.* Oxford University Press. New York.

Steinberg D (2004; 2005 a & b; 2006 a & b). *An interpretative history of the cholesterol controversy: Parts I - V*. Journal of Lipid Research. **45**, 1583-1593; **46**, 179-189; **46** 2037-2051; **47**, 1-14; **47**, 1339-1351.

Stilgoe J, Irwin A & Jones K (2006). *The received wisdom*. http://www.demos.co.uk/files/receivedwisdom.pdf

Sulik KK, Johnston MC & Webb MA (1981). *Fetal alcohol syndrome: Embryogenesis in a mouse model*. Science **214**, 936-938.

Susser M (1973). *Causal thinking in the health sciences: Concepts and strategies of epidemiology*. Oxford University Press. New York.

Susser E, Neugebauer R, Hoek HW, Brown AS, Lin S, Labovitz D & Gorman JM (1996). *Schizophrenia after prenatal famine: Further evidence*. Archives of General Psychiatry. **53**, 25-31.

Susser E, Schwartz S, Morabia A & Bromet EJ (2006). *Psychiatric epidemiology: Searching for the causes of mental disorders.* Oxford University Press. Oxford & New York.

Szabo R & Short RV (2000). *How does male circumcision protect against HIV infection?* British Medical Journal. **320**, 1592-1594.

Taubes G (1995). *Epidemiology faces its limits*. Science. **269**, 164-169.

Thistlewaite DL & Campbell DT (1960). *Regression-discontinuity analysis: An alternative to the ex post facto experiment*. Journal of Educational Psychology. **51**, 309-317.

Thornberry TP, Krohn MD, Lizotte AJ & Chard-Wiershem D (1993). *The role of juvenile gangs in facilitating delinquent behavior*. Journal of Research in Crime and Delinquency. **30**, 55-87.

Thorpe K, Rutter M & Greenwood R (2003). *Twins as a natural experiment to study the causes of mild language delay: II. Family interaction risk factors*. Journal of Child Psychology & Psychiatry. **44**, 342-355.

Tobin MD, Minelli C & Burton PR (2004). *Development of Mendelian randomization: From hypothesis test to 'Mendelian disconfounding'*. International Journal of Epidemiology. **33**, 21-25.

Tolmie JL (2002). Down syndrome and other autosomal trisomies. In Rimoin DL, Connor JM, Pyeritz RE & Korf BR eds. *Emery and Rimoin's principles of medical genetics, 4th Edition.* Churchill Livingstone. London & New York. 1129-1183.

Uchiyama T, Kurosawa M & Inaba Y (2007). *MMR Vaccine and regression in autism spectrum disorders: Negative results presented from Japan*. Journal of Autism and Developmental Disorders. **37**, 210-217.

Uher R & McGuffin P (2007). *The moderation by the serotonin transporter gene of environmental adversity in the aetiology of mental illness: review and methodological analysis*. Molecular Psychiatry. **1-16**, e-pub ahead of print.

Vandenbrouke JP (2004). *When are observational studies as credible as randomized controlled trials?* Lancet. **363**, 1728-1731.

Vandenbroucke JP, von Elm E, Altman DG, Gøtzche PC, Mulrow CD, Pocock SJ, Poole C, Schlesseman JJ & Egger M for the STROBE Initiative (2007) The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) Statement: Guidelines for Reporting Observational Studies. PLoS Medicines. **4(10)**, e297.

Vlajinac HD, Petrovic RR, Marinkovic JM, Sipetic SB, & Adanja BJ (1997). *Effect of caffeine intake during pregnancy on birth weight*. American Journal of Epidemiology. **145**, 335-338.

Wakefield AJ, Murch SH, Anthony A, Linnell J, Casson DM, Malik M, Berelowitz M, Dhillon AP, Thomson MA, Harvey P, Valentine A, Davies SE & Walker-Smith JA (1998). *Ileal-lymphoid-nodular hyperplasia, non-specific colitis, and pervasive developmental disorder in children*. Lancet. **351**, 637-641.

Wanless D (2002). *Securing our future health: Taking a Long Term View*. http://www.hmtreasury.gov.uk/consultations_and_legislation/wanless/consult_wanless_final.cfm

Weatherall D (2006). *The use of non-human primates in research.* An independent working group report sponsored by the Academy of Medical Sciences, Medical Research Council, Royal Society and Wellcome Trust. London.

Weller TH & Neva A (1962). *Propagation in tissue culture of cytopathic agents from patients with rubella-like illness*. Proceedings of the Society for Experimental Biology and Medicine. **111**, 215-225.

Wigle DT & Lanphear BP (2005). *Human health risks from low level environmental exposures: No apparent safety thresholds*. PLoS Medicine. **2**, e350.

Wilsdon J & Willis R (2004). *See through science. Why public engagement needs to move upstream*. http://www.demos.co.uk/files/Seethroughsciencefinal.pdf

Wilsdon J, Wynne B & Stilgoe J (2005). *The public value of science*. http://www.demos.co.uk/files/publicvalueofscience.pdf

Wilson PW, Garrison RJ, Castelli WP, Feinleib M, McNamara PM & Kannel WB (1980). *Prevalence of coronary heart disease in the Framingham Offspring Study: Role of lipoprotein cholesterols*. American Journal of Cardiology. **46**, 649-654.

Winship C, & Morgan SL (1999). *The estimation of causal effects from observational data*. Annual Review of Sociology. **25**, 659-707.

Women's Health Initiative Steering Committee (2004). *Effects of conjugated equine estrogen in postmenopausal women with hysterectomy. The Women's Health Initiative randomized controlled trial*. Journal of the American Medical Association. **291**, 1701-1712.

World Health Organization (2003). *WHO Framework Convention on Tobacco Control*. WHO. Geneva.

Wroe SJ, Pal S, Siddique D, Hyare H, Macfarlane E, Joiner S, Linehan JM, Brandner S, Wadworth JDF, Hewitt P & Collinge J (2006). *Clinical presentation and pre-mortem diagnosis of variant Creutzfeld-Jakob disease associated with blood transfusion: A case report*. Lancet. **368**, 2061-2067.

Zammit S, Allebeck P, Andreasson S, Lundberg I & Lewis G (2002). *Self reported use as a risk factor for schizophrenia in Swedish conscripts of 1969: Historical cohort study*. British Medical Journal. **325**, 1199-1203.

Zhou W, Liu G, Miller DP, Thurston SW, Xu LL, Wain JC, Lynch TJ, Su L & Christiani DC (2003). *Polymorphisms in the DNA repair genes XRCC1 and ERCC2, smoking and lung cancer risk*. Cancer Epidemiology, Biomarkers and Prevention. **12**, 359-365.

Zoccolillo M, Pickles A, Quinton D & Rutter M (1992). *The outcome of childhood conduct disorder: Implications for defining adult personality disorder and conduct disorder*. Psychological Medicine. **22**, 971-986.

Academy of Medical Sciences
10 Carlton House Terrace
London, SW1Y 5AH

Tel: +44(0)20 7969 5288
Fax: +44(0)20 7969 5298

Email: info@acmedsci.ac.uk
Web: www.acmedsci.ac.uk